DOCUMENT RESUME

ED 147 358 TM 006 810

AUTHOR Dyer, Robert F.; And Others
TITLE Questionnaire Construction Manual.
INSTITUTION Army Research Inst. for the Behavioral and Social Sciences, Ft. Hood, Tex.
REPORT NO ARI-P-77-1
PUB DATE Jul 76
CONTRACT DAHC-19-74-C-0032
NOTE 197p.; For related document, see TM 006 811

EDRS PRICE MF-$0.83 HC-$10.03 Plus Postage.
DESCRIPTORS Data Analysis; *Interviews; *Manuals; Pretesting; *Questionnaires; *Surveys; *Test Construction

ABSTRACT

This manual has been prepared primarily for the use and guidance of those who must develop or administer questionnaires as part of Army field tests and evaluations. The general content and concepts, however, should be useful to anyone involved in constructing or administering surveys, interviews, or questionnaires. Guidance on (1) preparing, assembling, and arranging items in questionnaires, (2) the importance of and procedures for pretesting, and (3) respondent characteristics that influence questionnaire results are included. The guide deals briefly with analysis and evaluation of responses, and discusses interview presentation. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED147358

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AD

TM

([illegible])

P-77-1

QUESTIONNAIRE CONSTRUCTION MANUAL

FORT HOOD FIELD UNIT

U. S. Army

Research Institute for the Behavioral and Social Sciences

July 1976

TM006 810

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the
Deputy Chief of Staff for Personnel

J. E. UHLANER Technical Director	W. C. MAUS COL, GS Commander

---

Research accomplished under
contract to the Department of the Army

Operations Research Associates

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 1300 Wilson Boulevard, Arlington, Virginia 22209.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
P-77-1

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
QUESTIONNAIRE CONSTRUCTION MANUAL

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
R. F. Dyer, J. J. Matthews, C. E. Wright, and
K. L. Yudowitch

8. CONTRACT OR GRANT NUMBER(s)
DAHC 19-74-C-0032

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Operations Research Associates
Palo Alto, California

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
2Q763731A775

11. CONTROLLING OFFICE NAME AND ADDRESS
TRADOC Combined Arms Test Activity
Fort Hood, Texas 76544

12. REPORT DATE
July 1976

13. NUMBER OF PAGES
183

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
U.S. Army Research Institute for the Behavioral and Social Sciences
ARI Field Unit-Fort Hood HQ TCATA (PERI-OH)
Fort Hood, Texas 76544

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Contracting Officer's Technical Representative was George M. Gividen, Chief, ARI Field Office at Fort Hood, Texas. Companion volume is "Questionnaire Construction Manual Annex: Literature Survey and Bibliography."

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Question construction
Test development
Item development
Questionnaire administration
Instructional manual on test construction

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This manual has been prepared primarily for the use and guidance of those who are tasked to develop or administer questionnaires as part of Army field tests and evaluations. The general content and concepts, however, should be useful to anyone involved in constructing or administering surveys, interviews, or questionnaires. Chapters 2-10 present guidance on preparing, assembling, and arranging items in questionnaires. Chapter 11 discusses the importance of and procedures for pretesting, and Chapter 12 gives respondent characteristics

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. that influence questionnaire results. Chapter 13 deals briefly with analysis and evaluation of responses, and Chapter 14 discusses interview presentation.

Army Project Number
2Q763731A775

TCATA
DAHC-19-74-C-0032

QUESTIONNAIRE CONSTRUCTION MANUAL

Dr. Robert F. Dyer
Ms. Josephine J. Matthews
Dr. Calvin E. Wright
Dr. Kenneth L. Yudowitch
Operations Research Associates

REVISED BY:

Dr. Charles O. Nystrom, ARI

Submitted by:
George M. Gividen, Chief
Fort Hood Field Unit

July 1976

Approved by:

Joseph Zeidner, Director
Organizations and Systems
Research Laboratory

J. E. Uhlaner, Technical Director
U.S. Army Research Institute for
the Behavioral and Social Sciences

TABLE OF CONTENTS

I. Introduction

- A. Purpose and Organization of This Manual
- B. Definition of Questionnaire
- C. Conventions Used in This Manual
- D. Keeping This Manual Up to Date
- E. Reporting Problems and Suggestions for Improvement

II. Major Questionnaire Types and Administration Procedures

- A. Overview
- B. Types of Questionnaires Discussed in This Manual
- C. Ways That Questionnaires Can Be Administered
- D. Structured Interviews Versus Other Types of Questionnaires

III. Content of Questionnaire Items

- A. Overview
- B. Determining Questionnaire Content Preliminary Research
- C. Other Considerations Related to Questionnaire Content

IV. Types of Questionnaire Items

- A. Overview
- B. Open-Ended Items
- C. Multiple Choice Items
- D. Rating Scale Items
- E. Ranking Items
- F. Forced Choice Items
- G. Card Sorting Items/Tasks
- H. Semantic Differential Items
- I. Other Types of Items

V. Attitude Scales and Scaling Techniques

- A. Overview
- B. Thurstone Scales
- C. Likert Scales
- D. Guttman Scales
- E. Other Scaling Techniques

VI. Preparation of Questionnaire Items

A. Overview
B. Mode of Items
C. Wording of Items
D. Difficulty of Items
E. Length of Question/Stem
F. Order of Question/Stems
G. Number of Response Alternatives
H. Order of Response Alternatives

VII. Response Anchoring

A. Overview
B. Types of Response Anchors
C. Anchored Versus Unanchored Scales
D. Amount of Verbal Anchoring
E. Procedures for the Selection of Verbal Scale Anchors
F. Scale Balance, Midpoints, and Polarity

VIII. Empirical Bases for Selecting Modifiers for Response Alternatives

A. Overview
B. General Considerations in the Selection of Response Alternatives
C. Selection of Response Alternatives Denoting Degrees of Frequency
D. Selection of Response Alternatives Using Order of Merit Lists of Descriptor Terms
E. Selection of Response Alternatives Using Scale Values and Standard Deviations
F. Sample Sets of Response Alternatives

IX. Physical Characteristics of Questionnaires

A. Overview
B. Location of Response Alternatives Relative to the Stem
C. Questionnaire Length
D. Questionnaire Format Considerations
E. Use of Answer Sheets

X. Considerations Related to Questionnaire Administration

A. Overview
B. Instructions
C. Anonymity for Respondents
D. Motivational Factors
E. Administration Time
F. Characteristics of Administrators

G. Administration Conditions
H. Training of Field Test Evaluators
I. Other Factors Related to Questionnaire Administration

XI. Pretesting of Questionnaires

A. Overview
B. Guidelines for Pretesting Questionnaires

XII. Characteristics of Respondents That Influence Questionnaire Results

A. Overview
B. Social Desirability and Acquiescence Response Sets
C. Other Response Sets or Errors
D. Effects of General Pretest Attitudes of Respondents
E. Effects of Demographic Characteristics of Responses

XIII. Evaluating Questionnaire Results

A. Overview
B. Scoring Questionnaire Responses
C. Data Analyses

XIV. Interview Considerations

A. Overview
B. Structured and Unstructured Interviews
C. Interviewer's Characteristics Relative to Interviewee
D. Situational Factors
E. Training Interviewers
F. Data Recording and Reduction
G. Special Interviewer Problems

ANNEX. Literature Survey and Bibliography

LIST OF FIGURES

Figure	Title	Section	Page
IV-B-1	Examples of Open-Ended Items	IV-B	1
IV-C-1	Examples of Multiple Choice Items	IV-C	2
IV-D-1	Examples of Numerical Rating Scale Items	IV-D	1
IV-D-2	Examples of Graphic Rating Scale Item	IV-D	2
IV-E-1	Examples of Ranking Items	IV-E	2
IV-F-1	Examples of Forced Choice Items	IV-F	2
IV-H-1	Examples of Semantic Differential Items	IV-H	2
IV-I-1	Examples of Check Lists	IV-I	1
IV-I-2	Examples of Formats Providing for Supplementary Responses	IV-I	3
VI-C-1	Examples of Question Form and Incomplete Statement Form of Stem	IV-C	2
VI-C-2	An Insufficiently Detailed Question Stem, Plus Revision	VI-C	4
VI-C-3	Examples of Loaded Questions	VI-C	7
VI-C-4	Examples of Leading Questions	VI-C	8
VI-C-5	Example of a Question Asking the Respondent to Criticize	VI-C	9
VI-C-6	Examples of Double-Barreled Questions and Alternatives	VI-C	10
VI-C-7	Example of Ambiguous Question and Alternative	VI-C	11
VI-C-8	Alternate Ways of Expressing Directionality and Intensity	VI-C	13
VI-D-1	Example of Hard to Understand Item and Alternative	VI-D	1
VI-H-1	Example of Rating Scale Item with Alternate Response Alternatives Order	VI-H	2
VII-B-1	Types of Response Anchors	VII-B	2

Figure	Title	Section	Page
VII-F-1	Examples of Scale Balance, Midpoints, and Polarity	VII-F	2
VIII-B-1	Two Formats Using "Outstanding" and "Superior"	VIII-B	12
VIII-B-2	Response Alternatives Frequently Recommended by ARI	VIII-B	13
IX-B-1	Arrangement of Items With Same Rating Scale Response Alternatives	IX-B	2
X-C-1	A Second Example of a Privacy Act Statement	X-C	4

LIST OF TABLES

Table	Title	Section	Page
VIII-B-1	Words Considered Unrateable by Subjects	VIII-B	2
VIII-B-2	Words Exhibiting Bimodality of Response	VIII-B	3
VIII-B-3	Sample List of Phrases Denoting Degrees of Acceptability	VIII-B	5
VIII-B-4	A second Sample List of Phrases Denoting Degrees of Acceptability	VIII-B	5
VIII-B-5	Neutral Term Scale Values and Standard Deviations as Determined by Several Different Studies	VIII-B	7
VIII-C-1	Degrees of Frequency	VIII-C	1
VIII-D-1	Order of Merit of Selected Descriptive Terms	VIII-D	1
VII-D-2	Order of Merit of Descriptive Terms Using "Use" as a Descriptor	VIII-D	2
VIII-E-1	Acceptability Phrases	VIII-E	2
VIII-E-2	Degrees of Excellence: First Set	VIII-E	3
VIII-E-3	Degrees of Excellence: Second Set	VIII-E	4
VIII-E-4	Degrees of Like and Dislike	VIII-E	5
VIII-E-5	Degrees of Good and Poor	VIII-E	6
VIII-E-6	Degrees of Good and Bad	VIII-E	7
VIII-E-7	Degrees of Agree and Disagree	VIII-E	8
VIII-E-8	Degrees of More and Less	VIII-E	9
VIII-E-9	Degrees of Adequate and Inadequate	VIII-E	10
VIII-E-10	Degrees of Acceptable and Unacceptable	VIII-E	11
VIII-E-11	Comparison Phrases	VIII-E	13
VIII-E-12	Degrees of Satisfactory and Unsatisfactory	VIII-E	14
VIII-E-13	Degrees of Unsatisfactory	VIII-E	14

Table	Title	Section	Page
VIII-E-14	Degrees of Pleasant	VIII-E	15
VIII-E-15	Degrees of Agreeable	VIII-E	15
VIII-E-16	Degrees of Desirable	VIII-E	16
VIII-E-17	Degrees of Nice	VIII-E	16
VIII-E-18	Degrees of Adequate	VIII-E	17
VIII-E-19	Degrees of Ordinary	VIII-E	17
VIII-E-20	Degrees of Average	VIII-E	18
VIII-E-21	Degrees of Hesitation	VIII-E	18
VIII-E-22	Degrees of Inferior	VIII-E	19
VIII-E-23	Degrees of Poor	VIII-E	19
VIII-E-24	Descriptive Phrases	VIII-E	20
VIII-F-1	Sets of Response Alternatives Selected so Phrases are at Least One Standard Deviation Apart and Have Parallel Wording	VIII-F	2
VIII-F-2	Sets of Response Alternatives Selected so That Intervals Between Phrases are as Nearly Equal as Possible	VIII-F	4
VIII-F-3	Sets of Response Alternatives Selected from Lists Giving Scale Values Only	VIII-F	6
VIII-F-4	Sets of Response Alternatives Selected Using Order of Merit Lists of Descriptor Terms	VIII-F	7

Chapter I: Introduction

A. Purpose and Organization of This Manual

1. Purpose

This manual has been prepared primarily for the use and guidance of those who are tasked to develop and/or administer questionnaires as part of Army field tests and evaluations, such as those conducted at the TRADOC Combined Arms Test Activity (TCATA) and the Combat Developments Experimentation Command (CDEC). The general content and concepts, however, are applicable to a variety of situations. As such, the manual should prove useful to all individuals involved in the construction and administration of surveys, interviews or questionnaires.

2. Organization

Information and guidance relating to the preparation of items for questionnaires and for their assembly and arrangement into a complete questionnaire are presented in Chapters II through X. Chapter XI discusses the importance of, and procedures for, pretesting questionnaires prior to their regular administration. Chapter XII discusses characteristics of respondents that influence questionnaire results. The analysis and evaluation of responses to a questionnaire are briefly dealt with in Chapter XIII. Finally, a number of considerations regarding the presentation of questions by means of an interview are discussed in Chapter XIV.

B. Definition of Questionnaire

As used in this manual, the word "questionnaire" refers to an ordered arrangement of items (questions, in effect) intended to elicit the evaluations, judgments, comparisons, attitudes, beliefs, or opinions of personnel. The content and format of the items may vary widely. A visual mode of presenting the items is employed. In the past, this meant that the items were typed or printed on paper, but now items can also be presented by closed circuit television or on a cathode ray tube under the control of a computer program. If the items are first read by an interviewer and then given verbally to the respondent, the questionnaire may also be termed a "structured interview." Hence, questionnaires and interviews have some common properties. Questionnaire items used to be responded to by scribing words or marks with a pen or pencil, but this aspect too has been enlarged to include typed, punched, and verbal responses.

While questionnaires are "data collection forms," not all data collection forms are questionnaires. Those forms used by personnel to enter instrument readings or to record their counts or observations (e.g., time of first detection, number of targets correctly identified, number of rounds fired) are not directly addressed in this manual.

C. Conventions Used in This Manual

1. Identification Scheme Used

This manual has been prepared in outline form to facilitate cross-referencing and later updating. The identification scheme that is used employs Roman numerals, capital and small letters, and numbers in the sequence: I A 1 a (1) [1] [a]. The major divisions, I, II, III, IV, etc., are called chapters. All other subdivisions are called "sections," with sections starting with capital letters (A, B, etc.) called "major sections." You are now, for example, reading Section I-C 1. To facilitate later updating, references within the manual are to sections and not pages.

2. Pagination

Each major section of this manual (e.g., I-C) starts on a new page, and pages are numbered within each major section. For example, this is Section I-C Page 1, or the first page of Section I-C.

3. Page Update Date

Immediately under each page number is the date that the page was drafted or revised. When a page has been revised, the date of the immediately previous version is also given in parentheses with the letter "s" meaning "superseded." For example, if I-D Page 1 dated 1 Jul 76 is revised on 10 Oct 76, the page number on the revised page would appear as:

I-D Page 1
10 Oct 76
(s. 1 Jul 76)

4. Table and Figure Identification

Both tables and figures are numbered sequentially within a major section, with a hyphen before the table or figure number. Examples are: Table VIII-B-1, Table VIII-B-2, Figure VI-A-1.

D. Keeping This Manual Up to Date

1. Updated Pages Should be Inserted as Received

It is anticipated that sections of this manual will be periodically corrected, revised, or otherwise updated. New pages should be inserted as soon as they are received. This will not only keep the manual up to date, but will facilitate adding pages received at an even later date. Appropriate instructions covering which pages to add and delete will accompany distributed update pages. When it appears useful, a list will also be provided showing the page numbers and dates of all pages that should be in the manual at that time.

2. Request for Updates

To be placed on the distribution list to receive updates to this manual, write to:

Chief
ARI Field Unit-Fort Hood
HQ TCATA (PERI-OH)
Fort Hood, Texas 76544

E. Reporting Problems and Suggestions for Improvement

As previously noted, it is anticipated that this manual will periodically be updated to improve its utility. To report errors, problems, or suggestions, write to:

Chief
ARI Field Unit-Fort Hood
HQ TCATA (PERI-OH)
Fort Hood, Texas 76544

Chapter II: Major Questionnaire Types and Administration Procedures

A. Overview

This chapter briefly summarizes the different types of questionnaires discussed in this manual (Section II-B) and ways that questionnaires may be administered (Section II-C). Detailed guidelines regarding which one to use in a given situation are included in subsequent chapters. Issues to consider when deciding whether to use a structured interview of some other type of questionnaire are presented in Section II-D, which also notes that combinations of methods may be employed. It is concluded that both structured interviews and other types of questionnaires have their place, and both have limitations.

B. Types of Questionnaires Discussed in This Manual

There are a number of techniques of data collection that can be used to measure human attributes, attitudes, and behavior. Some of these methods are observation, personal and public records, specific performances, sociometry, interviews, questionnaires, rating scales, pictorial techniques, projective techniques, achievement testing, and psychological testing. For this manual, however, attention has been restricted to a more limited number of data collection techniques: certain paper-and-pencil types of instruments broadly classed as questionnaires as defined in Section I-A 2, and including only some of the techniques mentioned above. A distinction has also been made in this manual between open-ended questionnaire items and closed-ended items. Open-ended items are those which permit the respondent to express his opinions in his own words and to indicate any qualifications he wishes. Closed-ended items, on the other hand, utilize response alternatives, such as multiple choice or true-false. Structured interviews are included within the definition of questionnaires used, since typically an interview form is developed and used by an interviewer both for asking questions and recording responses, much like a self-administered questionnaire. On the other hand, the unstructured interview makes no use of structured data collection forms. The interviewer is permitted to discuss the subject matter as he sees fit with no particular order or sequence. Of course, other interviews fall somewhere between these two extremes. In any case, unstructured interviews, where no structured response forms are used, are not included within the definition of questionnaires used in this manual.

C. Ways That Questionnaires Can Be Administered

There are a number of respects in which questionnaire administration may vary. However, in the usual field test settings, the modal questionnaire administration situation involves paper-and-pencil materials with the author/test officer administering the questionnaire face-to-face with a group of test players or evaluators.

1. Group Versus Individual Administration

Given a printed questionnaire, calendar time is saved by group administration. The task of statistical analysis can be initiated with less delay than if one were waiting on a series of individual administrations. An important determinant of group vs. individual is the time at which people complete their participation in the test. Most often all participants are through at the same time. All would be available for questionnaire administration as soon as they could be brought to an appropriate place or places. Prompt group administration gives the same, short amount of time for forgetting about test events to those who become the respondents. If there is an administrator, his time is conserved directly in proportion to the number of respondents he has in each administrative session.

2. Author-Administered Questionnaires

When the test officer or administrator who is familiar with the content of the questionnaire and the test's pruposes/objectives can administer the questionnaire, some advantages can be gained. The administrator's instructions and appeals may increase the number of respondents having desirable motivation to complete the questionnaire giving appropriate consideration to each item. If one employs a self-administration procedure such as might occur in a mailed-out questionnaire or if a poorly prepared stand-in plays the role of administrator, then the respondents must derive their instructions and some of their motivation from printed instructions (or from the poorly prepared stand-in). More things usually can end up going wrong when questionnairees are self-administered than when they are administered by a test administrator.

3. Remote Administrations

From the test officer's point of view this refers to a questionnaire administration event that he cannot conduct because of its distance from him and/or other demands on his time. This dimension, remote versus face-to-face, is similar but not identical to the previously noted dimension, self-administered versus author administered.

To avoid the possible disadvantages of self-administered questionnaires, the test officer must be able to afford another administrator, train him in the knowledge and skills associated with effective administration, and transport him to the "remote" administration location. If multiple administrations having location or timing differences to preclude the same administrator handle them are required, it would appear that the chances are increased that more respondents will experience more "difficulties" in answering the questions.

4. Other Materiel Modes

While providing the respondent with a printed questionnaire form and a pencil to mark/write his responses in the most common questionnaire administration procedures in field evaluations, other presentation modes have been used. In a card-sorting procedure that has been used with individuals and groups, each respondent reads statements of candidate problems and then places the slip in one of "n" piles according to his judgement of the severity of the "problem". Rarer because of expense and logistics problems is the setting up of a computer terminal where each respondent enters (types in) answers to questions that are displayed on a cathode ray tube (or other computer display device).
Chaper XII presents many other considerations related to questionnaire administration.

D. Structured Interviews Versus Other Types of Questionnaires

1. Issues to Consider

When deciding whether to use a structured interview or another type of questionnaire, a number of issues should be considered.

Included are the following:

a. If a structured interview is used, there must be enough qualified interviewers to expeditiously process all interviewees. Sometimes there are only a few personnel to be interviewed, or there is plenty of time available for interviews, so only one or two interviewers will be necessary. In other situations maybe only an hour or so may be available per interviewee; in these cases a large number of qualified interviewers must be available.

b. In most cases, respondents have a greater tendency to answer open-ended questions in an interview than when response is by paper and pencil.

c. Paper-and-pencil questionnaires may be less expensive, more anonymous, and completed faster than the same number of interviews.

d. Respondents seem to be less likely to report unfavorable things in an interview than in an anonymous questionnaire. Typically, questionnaires are also more likely than interviews to produce self-revealing data.

e. Issues involving socially acceptable or unacceptable attitudes and behaviors will elicit more bias in interviewee's responses.

f. During interviews, respondents often have a tendency to try to support the norms that they assume the interviewer adheres to.

g. Interviewers with biases on the issues under discussion may reflect them in the content they record as well as in what they fail to record.

h. Although a structured interview using open-ended questions may produce more complete information than a typical questionnaire containing the same questions, empirical research seems to indicate that responses to the typical questionnaire are more reliable; i.e., more consistent.

2. Combinations of Methods

There are some situations where a combination of methods of questioning might be used:

a. An interview might be used to obtain information for designing a paper-and-pencil questionnaire.

b. Personal interviews or telephone interviews might be used for respondents who do not return questionniares administered remotely (such as mail questionnaires).

c. When respondents are unable to give complete information during an interview, they can be left a copy of a questionnaire to complete and mail in, so that the necessity for a call-back is eliminated.

3. Conclusion

Both structured interviews and other types of questionnaires appear to have their advantages and disadvantages. The choice of which to use may well depend upon costs, which are generally lower for the typical questionnaire. The typical questionnaire is apparently more reliable, while the structured interview may provide more unique and more abundant information. If the dimensions of a problem have not been explored before, the best compromise would appear to be to use the interview approach with open-ended items to uncover the dimensions, and follow this by the use of the paper-and-pencil questionnaire with closed-end items to obtain more specific information.

Chapter III: Content of Questionnaire Items

A. Overview

The recommended general steps in preparing a questionnaire include preliminary planning, determining the content of questionnaire items, selecting question forms, wording of questions, formulating the questionnaire, and pretesting. As part of preliminary planning, the information requrired has to be determined, as do procedures required for administration, sample size, location, frequency of administration, experiemental design of the field test, and analyses to be used. Selecting question forms is a function of the content of the questionnaire items and requires knowledge of types of questionnaire items and scaling techniques. The wording of questions is the most critical and most difficult step. Formulating the questionnaire includes formatting, sequencing of questions, consideration of data reduction and analysis techniques, determining basic data needed, and insuring adequate coverage of required field test data. Pretesting involves using a small but representative group to insure that all questions are understandable and unambiguous.

This chapter considers the content of questionnaire items. Methods for determining questionnaire content are discussed first, and then other considerations related to questionnaire content are presented. The other steps noted above are discussed in subsequent chapters.

B. Determining Questionnaire Content Preliminary Research

1. Preliminary Research

If you have the job of developing a questionnaire for a field test, there are several things that should be done before starting to write questionnaire items.

a. Learn the test's objectives. Read the Outline Test Plan in order to learn what it says the test's purpose, scope, and objectives are. All data collection effort, including questionnaire administration, should be consistent with and supportive of the test's objectives.

b. What performance measures are planned for the test? One may be fortunate enough to be involved with a test for which the Detailed Test Plan has to a large extent been written. Try to discover what performance measures/data are to be collected. If performance data is to be collected on some aspects of the functioning of the system to be tested, then it may not be necessary to assess these functions via questionnaire items.

c. Consult others and prior test plans and reports. Many tests at CDEC and TCATA (and elsewhere) follow-up, or are similar to, prior testing. As a consequence, information may be readily available regarding prior related or similar tests. Test files or the Technical Information Center may provide a source for obtaining test plans and reports on relevant prior tests conducted by Army field test/experimentation agencies.

2. Using Interviews to Determine Questionnaire Content

If one's degree of experience seems meager relative to the complexities of the evaluation problem, he may employ group and/or individual interviews to assist in determining questionnaire content. Preferably this would be done after taking the steps noted above. The less one knows about a subject, the less structure one can impose on an interview dealing with the subject.

a. Conducting an unstructured group interview. Personnel are needed who have relevant operating experience with the system to be tested/evaluated - or with a sufficiently similar system. Arrange a common meeting place and time with about five to seven of them. It would be advantageous to have a meeting place that was not cramped for space, had comfortable chairs, a comfortable temperature, and where all discussants were free from other sources of distraction (sights and sounds, mainly).

If the interviewer's age and rank are several steps above or below the age and rank of the members of a homogeneous group of discussants, try (before the meeting) to get a person who is their contemporary (peer) in age and rank to lead and coordinate the discussions. Why? Because a mismatch may inhibit their discussion or produce too much submissive, agreeing behavior on their part.

If notes are being taken or the discussion is being tape recorded one should be unobtrusive about it. Don't shove/ point a microphone at a person as he starts to speak. He may be inhibited by this, or he may become a "ham".

The first several minutes should be spent in establishing rapport with the group. The purpose of the session should be covered, introduction of group members made, and other warmup devices used. The objective is to motivate as many respondents to give comments as possible. In the remainder of the session any or all of the following information-eliciting devices could be used:

(1) Discuss samples of the control item--ask the general question: "What problems have you had with this piece of equipment or system?" Follow up with who, what, where, when and why. Attempt to maximize the number of potential or actual problems posed. Strive for clarification of problem ideas, but do not criticize the comments, even if they are redundant with a previous contribution by

(2) Ask: "What do you consider to be the most important features (characteristics, qualities, etc.) of this equipment or system when used in the field?" Strive to get a multitude of adjectives and phrases here (e.g. ease of operation, weight, durability, portability, etc.)

(3) Use the aided recall technique: "Can you remember where and when you have encountered problems with this system?" (e.g., at night; when it's damp, etc.).

The recorded comments should be categorized and arranged by frequency. For example, how many of the comments on system operation stressed failure considerations?

b. Conduct semistructured personal interviews. As a next step, or as an alternative step to the group interview, one may employ a small number of representative respondents in a person-to-person interview format. Information produced from the unstructured group interviews provides general guidance to the specific evaluative information desired.

In this method of interviewing, the interviewer is given only general instructions on the type of information desired. He is left free to ask the necessary direct questions to obtain this information, using the wording and the order that seems most appropriate in the context of each interview. These interviews, like the unstructured group sessions, are useful in obtaining a clearer understanding of problems, and in determining what areas (evaluation criteria) should be included on the final questionnaire.

The only structure to the semistructured interview comes from a set of question categories that must be raised sometime during the interview. Questions on system experience, positive and negative features, and problems in field use, for example, can be phrased in any manner or sequence. Probing questions of the type: "Why do you feel that way?", "What do you mean by that statement?", and "What other reasons do you have?" can be utilized until the interviewer is satisfied that he has the necessary information considering time limitations, data requirements, and the willingness and ability of the respondents to verbalize their views.

In the semistructured interview, the interviewer has some flexibility in formulating and asking questions. This technique can, therfore, be only as effective in obtaining complete, objective, and unbiased information as the interviewer is skilled in formulating and asking questions. Thus interviewers may have to be trained in using this technique.

c. Develop the questionnaire. The use of the unstructured and semi-structured interviews as discussed above should enable the formulation of a questionnaire to obtain evaluative information. These interviews will provide guidance to the formulation of a sound survey instrument in the following respects:

(1) A better understanding of the factors or criteria which make up the mental set of individuals in evaluating systems and equipment.

(2) Some idea of the range of favorable and unfavorable opinions toward the system for each factor.

(3) Tentative knowledge of individual and group differential opinions toward the system tested.

Therefore, before drafting the formal questionnaire, the researcher must have a feel for: question categories (e.g., problem areas, positive aspects); response categories (e.g., evaluative factors); and the type of system operations information which is needed (e.g., in evaluating a new helmet suspension system, does respondent wear eyeglasses?).

3. Using the Critical Incident Technique to Determine Questionnaire Content

The critical incident technique consists of a set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness either in solving practical problems or in developing broad psychological principles. The technique calls for collecting observed incidents of behavior that have special significance and meet systematically defined criteria. It can be of assistance, therefore, in helping to determine the content of items to be included in a questionnaire.

Although there are a number of variations in the critical incident technique, the basic procedure consists of collecting records of specific behaviors related to the topic of concern. The behaviors might be noted by observers, or individuals can be asked to recall and record past specific behaviors judged to provide significant or critical evidence related to the topic of concern. As appropriate, behaviors related both positively and negatively to the area of concern should be noted. The records of behavior that are collected can then be analyzed and used as a basis for determining questionnaire content.

One of the examples of the use of the critical incident technique reported by Flanagan in the article noted in Section III-B 3, had to do with a study of combat leadership in the United States Army Air Forces in 1944. It represented "the first large-scale, systematic effort to gather specific incidents of effective or ineffective behavior with respect to a designated activity. The instructions asked the combat veterans to report incidents observed by them that involved behavior which was especially helpful or inadequate in accomplishing the assigned mission. The statement finished with the request, 'Describe the officer's action. What did he do?' Several thousand incidents were collected in this way and analyzed to provide a relatively objective and factual definition of combat leadership. The resulting set of descriptive categories was called the 'critical requirements' of combat leadership" (p. 328).

For more information on the critical incident technique, see, for example, the following two sources:

a. Barnes, T. I. The critical incident technique. Sociology and Social Research, 1960, 44, 345-347.

b. Flanagan, J. C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

4. Using Impressions of a Topic to Determine Attitude Scale Content

When the questionnaire is an attitude scale, a useful method for selecting items for it is to ask a group of individuals to write six statements giving their impressions of a topic, such as Army pay. From these, some smaller number of statements can be selected that are readable, intelligible, and capable of classification. These statements can then be sorted into several categories, such as the status of the topic and its good and bad features.

C. Other Considerations Related to Questionnaire Content

This section discusses a number of topics related to questionnaire content: questions that should be asked related to questionnaire content; sources of bias in questionnaire construction; and characteristics of good questions that affect questionnaire content.

1. Questions That Should Be Asked Related to Questionnaire Content

Asking yourself the following five questions may lay the foundation for a far more valuable questionnaire than would otherwise be produced:

a. Who needs the information? Knowledge of who needs the information will provide a source in the event answers are needed to the following four questions.

b. What decisions will be made based on your information? This will tell in part why the information is needed. Depending on what decision is going to be made, some kinds of information will make a difference and should be collected, and other kinds will not.

Suppose, for example, information is to be collected as a part of a test comparing a new item of equipment with an old standard item. The nature of the decision to be made is clear enough. It will be either selection of the new equipment, or retention of the old with which it is being compared. The basis for the decision will usually also be clear, From the small development requirement (SDR) or qualitative materiel requirement (QMR) which led to the development of the item being tested. Analysis of the QMR will identify the qualitative requirements the new equipment must have, and will give the start needed to develop questions.

c. What facts will affect the decision? While this may be a difficult question to answer, trying to do so should identify items of information that should be sought with the questionnaire. It may also head off the collection of unnecessary information.

d. Whom are you asking? To get good information, not only must a good question be asked, but it must be asked of someone who has the answer. It would not, for example, be reasonable to ask support troops in a supply depot questions about combat operations.

e. What are the consequences of a wrong answer? While this basically is an administrative question, it has an important bearing on field questionnaire design. Clearly, if it makes little difference which of two alternatives are chosen, it makes little difference if the information is collected. On the other hand, if there is a chance that substantial dollar savings will result from the use of a more effective training technique, or that millions of dollars will be wasted by buying a new piece of equipment which is not better than the old, it is necessary to design tests very well, and ask the right questions with great care.

2. Sources of Bias in Questionnaire Construction

Two primary sources of bias in questionnaire construction that have been identified are investigator bias and question bias.

a. Investigator bias arises from: choice of subject matter; study design and procedure; unfair or loaded phrasing of questions; and interpretation and reporting of results. Sources of such biases include: the questionnaire developer's relationship with the client; his personal involvement in a particular theoretical position or research technique; and those personal traits attributable to class, race, or political ideology. To reduce the impact of such bias, questionnaire developers need to: be aware of the problems; seek critiques from independent sources; carefully review previously published related reports; and continue to pursue technical improvement in their investigations.

b. Four ways that have been suggested of minimizing question bias when asking opinion questions are: ask many questions on the same topic; determine by scale analysis whether questions ask the respondents about the same dimensions of opinion (see Chapter V); ask "How strongly do you feel about this?" after each opinion question, and relate the content of opinion to the intensity of feeling.

Chapter IV: Types of Questionnaire Items

A. Overview

This chapter discusses various types of questionnaire items: open-ended items (Section IV-B), multiple choice items (Section IV-C), rating scale items (Section IV-D), ranking items (Section IV-E), forced choice and paired comparison items (Section IV-F), card sorting items/tasks (Section IV-G), and semantic differential items (Section IV-H). For each of these major item types, definitions and examples are presented, advantages and disadvantages are noted, and recommendations regarding their use in Army field test evaluations are given. Other types of items are noted in Section IV-I: check lists, matching items, arrangement items, and formats providing for supplementary responses.

It may be noted that a number of ways have been utilized in the professional literature for differentiating and classifying item types. Which types are special cases of other types could be debated at length. Unanimous agreement with the definitions given in this manual cannot, therefore, be anticipated.

B. Open-Ended Items

1. Definition and Examples

Open-ended items are those which permit the respondent to express his answer to the questions in his own words, and to indicate any qualifications he wishes. They are like general questions asked in an unstructured interview. By contrast, in a closed-ended item, all the answers/choices/responses permitted are displayed, and the respondent needs only to check his preferred choice. Examples of open-ended items are shown in Figure IV-B-1.

Figure IV-B-1

Examples of Open-Ended Items

1. Describe any problems you experienced in moving through the test course while wearing the new PRC-99 radio harness.
____
____

2. The M16 rifle is:______________________________
____

3. What do you think of the AR-15 rifle sight?_____________
____

2. Advantages of Open-Ended Items

a. Open-ended items allow for the expression of middle opinions that closed-ended items with two choices would not.

b. Open-ended items allow for the expression of issues of concern that may not have been identified by the question writer.

c. Open-ended items provide unique information.

d. Open-ended items are very easy to ask. This is useful when the question writer either does not know, or is not certain about, the range of possible alternative answers.

e. With an open-ended question it is possible to find out what is salient to the respondent, what his frame of reference is, and how strongly he feels.

f. There are times when more valid answers may be obtained from open- than closed-ended items. For example, there may be a tendency for respondents to inflate yearly income figures. Providing response alternatives may result in an even greater inflation.

3. Disadvantages of Open-Ended Items

a. Open-ended items are time consuming for the respondent.

b. A respondent may say that he has no problem rather than take the time to write out what the problem is. Item 1 in Figure IV-B-1 is poor in this respect, but item 2 is worse.

c. Open-ended items often leave the respondent on his own to determine what is relevant in evaluation. For instance, item 2 in Figure IV-B-1 leaves the respondent to determine what is relevant in evaluating the M16 rifle. This is inappropriate; open-ended questions should not be used to bypass the understanding of operations that the questionnaire writer should have or acquire before he prepares the final version of the questionnaire.

d. Questionnaires that use closed-ended items are generally more reliable than those using open-ended items.

e. Open-ended questions, answered by motivated respondents, are capable of overloading data analysts. They usually cannot be handled by machine analysis methods without lengthy preliminary steps. Analysis of the responses to an open-ended question usually must be done by someone who has substantial knowledge about the question's content, rather than by a statistical clerk. They are often difficult to code for analyses. Thus the data analysis problem can grow into a major project unless some other form of question is used.

f. Open-ended questions may be easier to misinterpret since the respondent does not have a set of response alternatives available which might in themselves provide the proper frame of reference.

g. Much of the material obtained from an open-ended question may be repetitious or irrelevant.

h. Open-ended questions are subject to more interviewer variations than closed-ended questions.

i. Open-ended items are often harder for the respondent to answer than closed-ended questions. For example, a respondent when asked his annual income may have to struggle to come up with a relatively specific figure, whereas when response alternatives are presented he need only indicate one of a number of ranges of income.

4. Recommendations Regarding Use

a. Open-ended questions should be rarely used and, even then, such questions should sharply focus the respondent's attention and thereby reduce his writing burden.

b. Sometimes a good procedure is to use an open-ended question with a small number of respondents as a pretest, in order to find out what the range of alternatives is. It may then be possible to construct good closed-ended questions that will be faster to administer and easier to analyze.

c. Open-ended questions are most useful when there are too many possible responses to be listed or forseen; when it is important to measure the saliency of an issue to the respondent; or when a rapport-building device is needed in an interview.

d. It is sometimes useful to include an open-ended question or so along with closed-ended questions in order to obtain verbatim responses or comments that can be used to provide "flavor" of responses in a report.

C. Multiple Choice Items

1. Definition and Examples

In a multiple choice item, the respondent's task is to choose the appropriate or best answer from several given answers or options. As used here, multiple choice items include dichotomous or two-choice items as special cases. And, since the permitted answers are available for selection, the multiple choice items may also be termed a closed-ended item.

Examples of multiple choice items are shown in Figure IV-C-1. Items 3, 4, and 5 are dichotomous or two-way.

A comparison of true-false items with nondichotomous multiple choice items is made in Section VI-G, since they are issues related to the number of response alternatives.

2. Advantages of Multiple Choice Items

a. As seen in item 2 of Figure IV-C-1, the questionnaire writer may select different numbers of response alternatives depending upon his knowledge of the respondent's experience or depending upon his decision to allow or disallow respondents to "sit on the fence" by including a "no preference" alternative. (See Section VI-C for wording of items, and Section VI-G regarding the number of response alternatives to employ).

b. Dichotomous items are relatively easy to develop, and permit rapid analyses.

c. Multiple choice items are easily scored, which means that data analysis is a relatively inexpensive process requiring no special content expertise.

d. Multiple choice items require considerably less time per respondent answer than open-ended items.

e. Multiple choice items put all persons on the same footing when answering. That is, each person will be able to consider the same range of alternatives when choosing an answer.

f. Multiple choice items are easy to administer.

Figure IV-C-1

Examples of Multiple Choice Items

1. What do you consider the most important characteristic of a good helmet? (Check one)

   _____ Comfort

   _____ Stability

   _____ Utility for wash basin

   _____ Protection

   _____ Weight

2. Which do you prefer, the M16 or the M14 rifle? (Check one)

   _____ M14

   _____ M16

   _____ No preference

3. Were you able to fire effectively from the frontal parapet emplacement?

   _____ Yes _____ No

4. Which do you prefer, the ABC helmet or the XYZ helmet?

   _____ ABC helmet _____ XYZ helmet

5. The M16 is a better rifle than the M14.

   _____ True _____ False

6. What is your marital status?

   _____ Single

   _____ Married

   _____ Divorced

   _____ Other (e.g., separated, widowed, etc.)

3. Disadvantages of Multiple Choice Items

a. Dichotomous items force the respondent to make a choice even though he may feel there are no differences between the alternatives, or he does not know enough about either to validly choose one. Furthermore, he is not permitted to say how much better one alternative is than the other.

b. Two alternatives might not be enough for some types of questions. The question designer may oversimplify an issue by forcing it into two categories.

c. There may be a tendency for respondents to choose an answer on the basis of a response set. (See Chapter XII).

d. Unless care is taken in the construction of multiple choice items, the response alternatives may overlap.

e. The question maker has to know the full range of significant possible alternatives at the time the multiple choice question is formulated.

f. Mutliple choice items must be worded with very great care. Otherwise, the information obtained may not be valid.

g. With dichotomous items any slight language difficulty or misunderstanding of even one word could change the answer from one extreme to another.

4. Recommendations Regarding Use

a. For some purposes the dichotomous or two-way question may be an improvement over the open-ended question in that it provides for faster and more economical analysis of data. However, it requires more care in its development.

b. Generally speaking, dichotomous multiple choice questions should be avoided If used, they should probably be followed up to determine the reason for a given response.

c. Nondichotomous multiple choice items are popular and have wide utility. They are recommended for general use as appropriate.

D. Rating Scale Items

1. Definitions and Examples

Rating scale items are a variation of multiple choice items. They are a means of assigning a numerical value to a person's judgment about some object. They call for the assignment of objects either along an unbroken continuum or in ordered categories along the continuum. The end result is the attachment of numbers to those assignments. Ratings may be made concerning almost anything, including people, groups, ourselves, objects, and systems.

There are a number of different forms of rating scale items, only two of which are shown here. Figure IV-D-1 shows examples of "numerical" scales. In item 1 a sequence of defined numbers is provided for the respondent.

Figure IV-D-1

Examples of Numerical Rating Scale Items

1. The cleaning kit for the M16 rifle is

_____ 7 very easy to use.
_____ 6 quite easy to use.
_____ 5 fairly easy to use.
_____ 4 borderline
_____ 3 fairly difficult to use.
_____ 2 quite difficult to use.
_____ 1 very difficult to use.

2. How satisfied or dissatisfied are you with the type of furniture in the barracks?

_____ Very satisfied
_____ Satisfied
_____ Borderline
_____ Dissatisfied
_____ Very dissatisfied

3. The training that I have received at Fort Hood has been

_____ very challenging.
_____ challenging.
_____ borderline.
_____ unchallenging.
_____ very unchallenging.

He is to indicate which defined number best fits his judgment about the object to be rated. Sometimes, the numbers are not shown on the form used by the respondent (e.g., items 2 and 3). Instead, the respondent reports in terms of descriptive cues and the numbers are attached later during analysis. The numbers assigned are in an arithmetic sequence, such as 5, 4, 3, 2, 1, depending upon the number of response alternatives used. They are usually assigned arbitrarily unless the response alternatives have been scaled using one of the procedures described in Section V-B. The order of perceived favorableness of commonly used words and phrases is discussed in Chapter VIII.

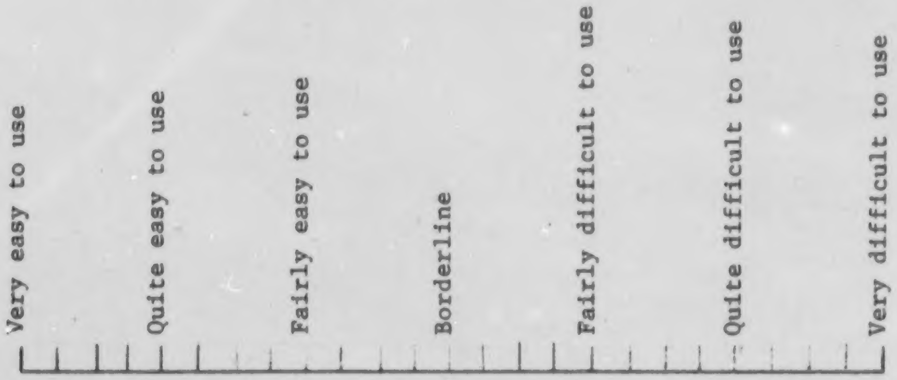
Figure IV-D-2 shows an example of a graphic rating scale. In the graphic scale, the descriptors are associated with points on a line or graph, and the respondent indicates his judgment by marking the point on the line which best fits his rating of the object. The line can be either horizontal or vertical. The graphic scale allows the respondent to place his judgment any place on the line, and thus he is not confined to discrete categories as he is with the numerical scale. It is, however, more difficult to score, but this can be facilitated with a stencil which divides the line into segments to which numbers are assigned.

The number of response alternatives to use is discussed in Section VI-G, the order of response alternatives in Section VI-H, and response anchoring in Chapter VII.

Figure IV-D-2

Example of Graphic Rating Scale Item

1. Place an X at the point on the scale that most clearly represents your opinion about the cleaning kit for the M16 rifle.



2. Advantages of Rating Scale Items

a. When properly constructed, the rating scale reflects both the direction and degree of attitude or opinion, and the results are amenable to analysis by conventional statistical tests (means, standard deviations, etc.).

b. Graphic rating scales allow for as fine a discrimination as the respondent is capable of giving, and the fineness of scoring can be as great as desired.

c. Rating scale items usually take less time to answer than do other type of items.

d. Rating scale items can be applied to almost anything.

e. Rating scale items are generally more reliable than two-way multiple choice items. They may be more reliable than paired comparisons items.

3. Disadvantages of Rating Scale Items

a. Rating scale items are more vulnerable to biases and errors than other types of items such as forced choice items.

b. Graphic rating scales are harder to score than other types of items.

c. The results obtained from the use of graphic rating scale items may imply a degree of precision/accuracy which is unwarranted.

4. Recommendations Regarding Use

The use of rating scale items is highly recommended for most questionnaires.

E. Ranking Items

1. Definition and Examples

Ranking items call for the respondent to indicate the relative ordering of the members of a presented group of objects on some presumably discriminable dimension, such as effectiveness, saltiness, overall merit, etc. By definition one does not have a scale by which the amount of difference between successive members is measured, nor is it implied in rank ordering that successive differences are even approximately equal. If respondents were being asked to give judgments on the size of intervals, the item would be something more than a ranking item.

Multiple choice items are so frequently used that one may inadvertently use this format when the ranking item format would provide more complete and reliable information. Item 1 in Figure IV-C-1 illustrates this point. Since a preponderance of respondents would check "protection" as a helmet's most important characteristic, only a small remainder of responses would be available as a basis for ordering the other characteristics. Some of the other characteristics might be achievable without sacrificing protection, so it would be desirable to have a reliable ordering of their importance.

As the number of objects to be ranked increases, the difficulty of assigning a different rank to each object increases even faster. This means that reliability (repeatability) is reduced. To counter this, one mya explicitly permit respondents to assign tied rankings to objects when the number of objects exceeds, say, 10 or more.

Examples of ranking items are shown in Figure IV-E-1.

2. Advantages of Ranking Items

a. The idea of ranking is familiar to respondents.

b. Ranking takes less time to administer, score, and code than paired comparisons items do, and there is some evidence that the results of the two have a linear relationship.

c. Ranking and rating techniques are generally comparable.

Figure IV-E-1

Examples of Ranking Items

1. Rank the following three methods of issuing starlight scopes to an infantry squad. Assign a "1" to the most effective, a "2" to the second most effective, etc. Do not assign tied rankings.

Ranking	Basis of Issue
_______	Scopes issued to AMG and SL
_______	Scopes issued to AMG, SL, and one rifleman
_______	Scopes issued to all squad members

2. How important are each of the following factors to you? Assign a "1" to the most important, "2" to the second most important, etc. Assign a different number to each of the four factors.

_______ Type of furniture in the barracks

_______ Army pay

_______ Medical service to soldiers

_______ Choice of duty station

3. Disadvantages of Ranking Items

a. Ranking items such as item 1 in Figure IV-E-1 do not reveal the respondent's judgment as to whether any of the objects are effective or ineffective in an absolute rather than just a relative sense. To learn this, another question must be asked.

b. Rank order items do not permit respondents to state the relative amounts of differences between alternatives.

c. The results from ranking items are open to question if the basis for ranking was not clear to the respondents.

d. Ranking is generally less precise than rating.

4. Recommendations Regarding Use

There are some situations where the intent of the questionnaire developer is best served with the use of one or more ranking items. Generally, however, rating scale items are probably preferable.

F. Forced Choice Items

1. Definition and Examples

It would appear that any multiple choice item could also be called a "forced choice" item because, afterall, the respondent is expected to choose one of the response alternatives. The instructions and/or the presence of an administrator put some degree of social pressure - social force - on the respondent. However, if a multiple choice item includes an "I don't know" response alternative, the pressure/force is almost totally removed. Likewise, on a rating scale item, the inclusion of a "neutral" or "borderline" response category allows the respondent to answer without committing himself.

So, for some questionnaire developers - in particular those who produce "forced choice self inventories" (see references) - a "forced choice" item strictly refers to one where the respondent must commit himself or herself. He may have to select one of a pair of choices, or two of three, or two of four. These three cases are illustrated in Figure IV-F-1.

2. Advantages of Forced Choice Items

a. Studies have indicated that the reliability and validities obtained from the use of forced choice items compare favorably with other methods.

b. Studies have also shown that forced choice items are more resistant than other items to the effects of bias.

c. The forced choice method has been used by a number of investigators in an attempt to control the tendency of individuals to answer self-report inventories in terms of response sets rather than giving "true" responses. (Response sets are discussed in Chapter XII.)

3. Disadvantages of Forced Choice Items

a. Respondents sometimes balk at picking unfavorable statements, or at being forced to make a choice.

b. Forced choice items take more time to develop than do other types of items.

c. Paired comparisons items where all phrases are paired take more time to administer, score, and code than do ranking items. Results from the two, however, may have a linear relationship.

Figure IV-F-1

Examples of Forced Choice Items

1. Check the one of the following two statements that is more Characteristic of what you like.

   _____ I like to travel.

   _____ I like to meet new people.

2. Check the one of the two following statements that is more characteristic of yourself.

   _____ I am honest.

   _____ I am intelligent.

3. Look at the following three activities. Mark an "M" by the one you like the most, and an "L" by the one you like the least.

   _____ Play baseball

   _____ Go to the craft shops

   _____ Attend boxing or wrestling matches

4. From the following four statements check the two that are most descriptive of your unit commander.

   _____ Serious-minded

   _____ Energetic

   _____ Very helpful

   _____ Gets along well with others

d. There is some question as to whether forced choice items overcome the biases or errors they are supposed to correct.

e. Some investigators have concluded that the generalization that self-report forced choice inventories are more valid than single stimulus forms of the same tests is not supported by a critical consideration of the relevant evidence.

Procedures for constructing forced choice items, and evaluative comments about them, can be found in a number of sources including the following:

a. Guilford, J. P. Psychometric methods (2nd ed.). New York: McGraw-Hill, 1954.

b. Nunally, J. C. Psychometric Theory. New York: McGraw-Hill, 1967, pp 484-485.

c. Sisson, E. D. Forced choice--the new Army rating. Personnel Psychology, 1948, 1, 365-381.

4. Recommendations Regarding Use

When test participants are deliberately given relevant experience with the operation of a weapons system, vehicle, or other system, the "I don't know" response alternative should normally be deleted from items that seek the participants' evaluations of the system.

G. Card Sorting Items/Tasks

1. Definition

With card sorting items/tasks, the respondent is given a large number of statements (e.g., 75), each on a slip of paper or card. He is asked to sort them into, say, nine or eleven piles. The piles are in rank order from "most favorable" to "least favorable" or "most descriptive" to "least descriptive", etc., depending upon the dimension to be used. Each pile usually is to have a specified number of statements placed into it as required to form a rough normal distribution. However, some investigators have argued that forcing a given distribution is not necessary. Ordinarily each pile is given a score value which is then assigned to the statements placed into it.

An extensive discussion of the use of card sorts (or, more generally, Q-technique and its methodology) appears in: Stephenson, W. The study of behavior. Chicago: University of Chicago Press, 1953.

2. Advantages of Card Sorting Items/Tasks

a. Card sorts appear to be capable of counteracting at least some of the biasing effects of response sets. (Response sets are discussed in Chapter XII.)

b. Some investigators believe that card sorting is a fast and interesting method of obtaining valid and reliable interview data.

c. With card sorts the respondent can shift items back and forth if he wishes to do so.

d. The card sort has greatest value when a comprehensive description of a single individual is desired.

e. Card sorts also have value for obtaining complex descriptions which can be compared systematically.

f. They can be used to obtain rating information on any issue.

3. Disadvantages of Card Sorting Items/Tasks

a. Card sorting items/tasks may take more time to construct than other types of items, and they generally take more time to administer and score.

b. Card sorts are more involved to administer than other types of questionnaire items.

4. Recommendations Regarding Use

Some authors think that card sorting is the method of choice if testing time is available. Its greatest value seems to be its ability to provide a comprehensive description of a single individual, or to obtain complex descriptions which can be systematically compared. Since it is more awkward to administer and score than other types of items, its use in Army field test evaluations is limited.

H. Semantic Differential Items

1. Definition and Examples

The semantic differential technique was initially developed as a general method of measuring meaning, and with it the meaning of a particular concept to a particular individual can be specified quantitatively. The technique has also been used to measure attitudes and values, particularly in the marketing area. In using the technique, the respondent is presented with a number of bipolar rating scales, usually but not always with seven points. The extreme of each scale is defined by an adjective. The respondent is given a set of such scales and is asked to rate each of a number of objects or concepts on every scale. To aid in interpretation, some coding scale can be used, usually numbers in a direct numerical sequence such as 1 through 7. Other more extensive scoring can be used, and results can be factor analyzed to search for the basic dimensions of meaning. However, the usefulness of the semantic differential as a research tool stems from the ability of the procedure to probe into both the content and the relative intensity of respondents' attitudes.

Examples of semantic differential items are given in Figure IV-H-1. A recommended text on the semantic differential is Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. The measurement of meaning. Urbana, Ill., University of Illinois Press, 1957. Norms have been collected on 20 scales for 360 words. They are reported in Jenkins, J. J., Russell, W. A., & Suci, J. An atlas of semantic profiles for 360 words. American Journal of Psychology, 1958, 71, 688-699.

2. Advantages of Semantic Differential Items

a. Evidence on the validity, reliability, and sensitivity of the scales has been offered.

b. Using some adjectives that do not seem appropriate to the concept under investigation may uncover aspects that reflect an attitude or feeling tone even though the respondent cannot put it into words.

c. Semantic differential items can be used to study the relative similarity of different concepts to the respondent, and to study changes over time.

d. Semantic differential items are relatively easy to construct, administer, and score.

Figure IV-H-1

Examples of Semantic Differential Items

1. Place an X in each of the following rows to describe your feelings about the M16 rifle.

Reliable ____: ____: ____: ____: ____: ____: ____ Unreliable

Heavy ____: ____: ____: ____: ____: ____: ____ Light

Good ____: ____: ____: ____: ____: ____: ____ Bad

Slow ____: ____: ____: ____: ____: ____: ____ Fast

Adequate ____: ____: ____: ____: ____: ____: ____ Inadequate

2. Place an X in each of the following rows to describe your feelings about the ABC helmet.

Reliable ____: ____: ____: ____: ____: ____: ____ Unreliable

Heavy ____: ____: ____: ____: ____: ____: ____ Light

Good ____: ____: ____: ____: ____: ____: ____ Bad

Slow ____: ____: ____: ____: ____: ____: ____ Fast

Adequate ____: ____: ____: ____: ____: ____: ____ Inadequate

3. Disadvantages of Semantic Differential Items

   a. If care is not taken, the two adjectives chosen for the extremes will not define some kind of scale or dimension between them.

   b. The value of semantic differential items depends on the suitable choice of the bipolar adjectives and concepts.

   c. There is a potential response error present in the respondents' interpretations of the meaning of the polar descriptions. However, there appears to be a balancing out over a number of administrations.

   d. The semantic differential is complex to score and analyze using the traditional procedures.

4. Recommendations Regarding Use

There are a number of investigators that advocate the use of the semantic differential. Others, however, have questioned whether it may be a rather complicated way of developing a measure that is more readily and reliably secured by other means. It is reasonable to assume that the technique could easily be expanded to identify attitudes and the intensity of the attitudes toward the attractiveness of a particular military specialty, the capacities of a specific piece of equipment to perform, or any other characteristic set which can be described by bipolar adjectives. However, since the analysis of sets of semantic differential items is somewhat involved, the technique has not been widely used for routine Army field test evaluations.

I. Other Types of Items

1. Check Lists

Check lists are instruments in which responses are made by checking the appropriate statement or statements in a list of statements. Examples are shown in Figure IV-I-1.

Figure IV-I-1

Examples of Check Lists

1. Which of the following are important to consider when deciding whether or not to make a career of the Army? Check all that apply.

_____ Leadership of NCO's

_____ Opportunity for promotion

_____ Playboy magazines in the Post Exchange

_____ Latrine in crafts shops

_____ Army pay

_____ Choice of duty stations

_____ Civilian opinion of Army

_____ Reenlistment bonuses

_____ Hours of work in a work week

2. Please check all the characteristics which Backpack A possess.

_____ Durability

_____ Lightness

_____ Wearing comfort

_____ Assessibility of items

_____ Ease of putting on and taking off

_____ Other (specify:) ______________________________

Compared to rating scales, which give a numerical value to some sort of judgment, check lists are relatively crude. They are, however, quite useful when rating information is not needed or when information is needed regarding which of a number of attitudes are significant to a respondent. Other issues regarding the use of check lists are as follows:

a. Check lists should use terms like the respondent uses.

b. Response set can be somewhat controlled if the respondent is asked to check a stated number of items, or if upper or lower limits are set.

c. There is some evidence that a higher rate of claim or assertion is obtained from check lists than from open-ended items.

d. It is usually not known if check lists cover the appropriate attributes.

e. Adjective check lists are sometimes used, especially to elicit stereotypes about people or nations. They are similar to rating scales.

2. Matching Items

With matching items, the respondent is given two columns of items and is asked to pair each item in the first column with an associated item in the second. In general, it is not desirable to have the same number of items in each column. Both sets of items should constitute a homogeneous set, and any item in the second column should look like it could go with any item in the first column.

Matching items are best used in achievement testing. Since they have little utility in Army field test evaluations, they are not discussed in greater detail.

3. Arrangement Items

With an arrangement item, a number of statements are presented in random order, and the respondent arranges them in a given way. For example, steps in a sequence of events or procedures may be rearranged in order of occurence or performance. Or, causes may be rearranged in order of importance in bringing about a certain effect.

There may be some situations where arrangement items may be useful in Army field test evaluations; however, the scoring of the items is difficult. The use of such items is, therefore, extremely limited.

4. Formats Providing for Supplementary Responses

The questionnaire writer is not limited to the major item formats described in this chapter. Formats providing for supplementary responses can also be used. Examples are shown in Figure IV-I-2.

Figure IV-I-2

Examples of Formats Providing for Supplementary Responses

1. The starlight scope is able to detect aggressor movements:

_____ very effectively.

_____ effectively.

_____ borderline.

_____ ineffectively.

_____ very ineffectively.

Explain: ____

____

2. What style of leadership was used by the most effective squad leader you served under? (Check one)

_____ democratic and friendly

_____ friendly with most; authoritarian with the others

_____ sometimes authoritarian; sometimes acts like one of the men

_____ usually authoritarian; avoided making close friends

_____ other (please describe) ____________________________

____

Notice that the extra response alternative in Example 2 allows the respondent in effect to make an open-ended item out of a multiple choice item. Few test respondents, however, elect to do this. Inclusion of the supplementary or write-in option commits you to extra data reduction and analysis effort that would have been unnecessary had you anticipated and included all reasonable response alternatives.

Chapter V: Attitude Scales and Scaling Techniques

A. Overview

At times, the questionnaire developer will wish to treat the total group of items on a questionnaire as a single measuring scale, and from them obtain a single overall score on whatever he is interested in measuring. This is a common practice, especially with the measurement of attitudes. A typical attitude scale is composed of a number of questions/statements selected and put together from a much larger number of questions/statements according to certain statistical procedures. Some of these procedures, called scaling techniques, are discussed in this chapter.

A distinction is needed, however, between two ways in which the term scale is used in this manual. An attitude scale could be constituted of items each one of which employs a response scale. Aspects of response scales are discussed in Chapter VII on "Response Anchoring." A component of score could be achieved on each item. Adding these item scores together - which means considering the whole set of items as a scale - produces a total attitude score for the individual respondent.

There are, generally speaking, two general methods for the construction of scales such as attitude scales. The first method makes use of a judging group and one of the psychological scaling methods developed by Thurstone, as discussed in Section V-B. It results in a set of statements being assigned scale values on a psychological continuum. The continuum may be favorableness, unfavorableness, like-dislike, or any other judgment. The psychological scaling methods, therefore, have considerably greater application than for the scaling of attitudes. They can be used to scale statements or objects. They have been used, for example, to determine the perceived favorableness of words and phrases commonly used as rating scale response alternatives, as discussed in Chapter VIII.

The second general method is based on the direct responses of agreement or disagreement with attitude statements and does not result in a set of statements being assigned scale values on a psychological continuum. Both the Likert and Guttman scales discussed in Sections V-C and V-D are examples of this latter method.

For information (relating to attitude scaling and scaling techniques) beyond that contained in this manual the following references may be consulted.

1. Edwards, A. L. Techniques of attitude scale construction. New York: Appleton-Century-Crofts, 1957.

2. Guilford, J. P. Psychometric methods (2nd ed.). New York: McGraw-Hill, 1954.

3. Gulliksen, H., & Messick, S. (Eds.). Psychological scaling: Theory and applications. New York: John Wiley, 1969.

4. Lemon, N. Attitudes and their measurement. New York: John Wiley, 1974.

5. Nunnally, J. C. Psychometric Theory. New York: McGraw-Hill, 1967.

6. Thurstone, L. L. The measurement of values. Chicago: University of Chicago Press, 1959.

7. Torgerson, W. S. Theory and methods of scaling. New York: John Wiley, 1958.

B. Thurstone Scales

This section discusses three scaling methods developed by L. L. Thurstone. For additional detail, see the texts referred to in Section V-A.

1. Method of Equal Appearing Intervals

Thurstone's method of equal appearing intervals was the first major method of attitude scaling to be developed. It was assumed that a group of statements of opinion about a particular issue could be ordered on a continuum of favorableness, unfavorableness, and that the ordering could be such that there appears to be an equal distance between the adjacent statements on the continuum.

The following steps are followed in the method of equal appearing intervals:

a. From the literature or pilot interviews, a large number of statements (100 to 200) are compiled about the attribute or object of an attitude under study. Irrelevant, ambiguous, or poorly worded statements would not be selected.

b. A number of judges, at least 50, are obtained. They should be similar to those individuals who will respond to the final statements on the questinnaire. The judges independently sort each statement into one of 11 piles. The first pile is defined as "Unfavorable" or "Most unfavorable," the middle or sixth pile is defined as "Neutral," and the eleventh pile is defined as "Favorable" or "Most favorable." The other files are left undefined. The judges are told that the intervals between piles or categories are to be regarded as subjectively equal. They are also instructed to ignore their own agreement or disagreement with each item, and to judge each item in terms of its degree of favorableness-unfavorableness.

c. The scale value for each item is usually determined by computing its mean or median, over all judges.

d. Twenty to 25 statements with little dispersion in their scale values are then selected for use. The statements are selected so that the intervals between statements' scale values are approximately equal and/or are relatively equally spaced on the psychological continuum.

e. The finally selected statements are usually placed in random order for presentation to respondents. The respondent is asked to indicate which statements he agrees with, and which he disagrees with.

f. The respondent's score is the mean or median scale value of those statements for which he marked "Agree."

Some considerations for use of the Equal Appearing Intervals method are:

a. The method of equal appearing intervals is designed to provide an interval scale as its output. The scale is at least ordinal (ranked).

b. The method is useful when there are a large number of statements involved.

c. Scale values from widely differing groups of judges appear to correlate highly with one another so long as judges with extreme views are eliminated.

d. Graphic or numerical rating scales can be used by the judges instead of having the statements sorted into piles. Though 11 categories are usually used, some other number can be employed.

2. The Method of Paired Comparisons

Thurstone developed a procedure for deriving an interval scale based upon what has been called the Law of Comparative Judgment. Basically, it is a method by which statements such as "A is stronger than B," "B is stronger than C," etc., are used to provide a scale with interval properties. The objects or statements to be ranked are presented two at a time, and the respondent is asked to choose between them. All possible combinations of pairs have to be presented. Hence the procedure becomes very cumbersome when there are more than 15 or so items. The determination of scale values is also laborious. Since the procedure is not used much in applied research, additional detail is not presented here.

3. The Method of Successive Intervals

The method of successive intervals is similar to the method of equal appearing intervals. However, no assumption is made concerning the psychological equality of the category intervals.

It is only assumed that the categories are in correct rank order and that their boundary lines are relatively stable. The procedure involves estimating the widths of the categories along the psychological continuum, and, from these reference points, the scale values of the statements can be obtained. Research has shown that there is a linear relationship between scales constructed by the method of paired comparisons and by the method of successive intervals.

C. Likert Scales

The Likert method of scale construction was developed because the Thurstone procedures require extensive work and make assumptions regarding the independence of item statements. The Likert method assumes that all statements reflect the same attitude dimension and are hence related to each other. The Likert approach does not assume equal intervals between the scale values. It is sometimes called the method of summated ratings.

The steps in Likert scale construction are as follows:

1. Statements are classified in advance as "Favorable" or "Unfavorable." No attempt is made to find an equal distribution of statements over the whole range of the attitude of concern, and no attempt is made to scale the statements.

2. A pretest is then conducted. In the pretest the respondents indicate their degree of agreement with every statement, usually using five response alternatives: strongly agree, agree, undecided, disagree, and strongly disagree.

3. Each descriptor is assigned a numerical weight (e.g., +2, +1, 0, -1, -2) usually based on a given series of integers in arithmetical sequence.

4. Each respondent is assigned a score that represents the algebraic summation of weights associated with each item checked. In the scoring process weights are assigned such that the direction of attitude, favorable to unfavorable, is consistent over items. For example, if an +2 is assigned to "Strongly agree" for favorable statements, a -2 should be assigned to "Strongly agree" for unfavorable statements.

5. The statements finally selected for use in the questionnaire are those which appear to discriminate best between respondents with the highest and lowest total scores. Usually about half of the statements are favorable, half unfavorable.

6. In the final questionnaire, a score is obtained by summing the numerical weights assigned to the

Factors to be taken into consideration when deciding whether to use Likert scales include:

1. Likert scales take less time to construct than Thurstone scales.

2. It is possible to construct scales by the Likert and Thurstone methods which will yield comparable scores.

3. Likert scales have only ordinal properties. If there is a large dispersion about a respondent's mean score, however, even those properties have limited meaning. If the sole purpose of a scaling procedure is to rank respondents according to the degree to which they hold some attitude, then Likert scales are efficient because of their ease of administration.

4. In addition to lacking metric properties, Likert summated scores lack a neutral point. The interpretation of a score cannot be made independently of the distribution of scores of some defined group. However, percentile or deviation-type norms can be calculated if the sample size is large enough.

5. For the same number of items, scores from Likert scales may be more reliable than scores from Thurstone scales.

D. Guttman Scales

Guttman's approach to scaling is called scalogram or scale analysis. It is a deterministic model; it considers its scales are close to being rulers-measures of length. The essence of the method is to determine whether a series of statements can be appropriately scaled. An attempt is made to identify a set of statements which actually reflect a unidimensional scale and have a cumulative nature. When the goal is achieved, two or more persons receiving the same score will have responded in the same way to all of the statements.

As an example, the following four questions comprise a Guttman scales:

	Yes	No
a. The United Nations is mankind's savior	___	___
b. The United Nations is our best hope for peace	___	___
c. The United Nations is a constructive force in the world	___	___
d. We should continue our participation in the United Nations	___	___

The expected pattern of responses to these questions is "triangular".

	Person			
Item	1	2	3	4
a	x			
b	x	x		
c	x	x	x	
d	x	x	x	x

This means that, for any person who answers yes to item "a", there is a high probability that he will answer yes to the other items. A person who says no to "a" but yes to "b" has a high probability of answering yes to the other items, and so on.

The major steps in scalogram analysis are too complex to summarize here, but are found in some of the references in Section V-A. Procedures are available for:

1. Measuring the amount of error due to imperfect scalability.

2. Ordering the statements so that the response patterns provide the least amount of error.

3. Deterimining the extent to which the data approximate the perfect case.

4. Improving the scalability of the statements via category combinations, statement discarding, etc.

There have been many critics of scalogram analysis. Some feel that there is no really effective way of selecting good items by this approach. However, the procedure is considered useful if one is concerned with unidimensionality or if one wishes to examine small changes in attitudes. It is, however, laborious. No instances of past use in field testing situations are known.

E. Other Scaling Techniques

Numerous other scaling techniques and combinations of methods are reported in the literature. A discussion of them is, however, outside the current scope of this manual.

Chapter VI: Preparation of Questionnaire Items

A. Overview

Once a decision has been made regarding the type or types of items that are to be used in a questionnaire (see Chapter IV), attention must be given to the actual development of the items. This chapter, then, addresses the following topics: mode of questionnaire items; wording of items for both question stems and response alternatives; difficulty of items; length of question stem; order of question stem; number of response alternatives and order of response alternatives. The related topic of response anchoring is considered in Chapter VII.

As used in this manual, a distinction has been made between a questionnaire item, a question stem, and response alternatives. A questionnaire item has both a question stem and response alternatives. The response alternatives are the answer choices for the question. (They are sometimes called "options.") The question stem is that part of the item that comes before the response alternatives.

B. Mode of Items

Questionnaire items are usually presented to a respondent in printed form. However, it is possible to present items or stimuli pictorially. There is some evidence that there are no significant differences in subjects' responses to verbal and pictorial formats. Using a pictorial format may facilitate obtaining responses from respondents with limited verbal comprehension, who might have difficulty responding to questions employing lengthy definitions of concepts or objects. If pictures are used, they should be pretested for clarity of their presentation of the concept or object to be evaluated.

In cases where it is known that the respondents have very low reading ability, it may be desirable to present the questionnaire orally. A tape player-recorder may be used for this purpose also.

C. Wording of Items

The wording of questionnaire items is a critical consideration in obtaining valid, relevant, and reliable responses. Consider, for example, the following three questions that were administered by Payne (see reference below) to three matched groups of respondents:

a. "Do you think anything should be done to make it easier for people to pay doctor or hospital bills?"

b. "Do you think anything could be done to make it easier for people to pay doctor or hospital bills?"

c. "Do you think anything might be done to make it easier for people to pay doctor or hospital bills?"

These questions differed only in the use of the words "should," "could," or "might," terms that are often used as synonyms even though they have different connotations. The percent of "Yes" replies to the questions were 82, 77, and 63, respectively. The difference of 19% between the extremes is probably enough to alter the conclusions of most studies.

A number of matters related to the wording of questionnaire items are considered in this section. Some of the suggestions made are based upon experimental research. Others are based upon experience, intuition, and common sense. Several sources offering principles of question wording are:

a. Roslow, S., & Blankenship, A. B. Phrasing the question in consumer research. Journal of Applied Psychology, 1939, 23, 612-622.

b. Jenkins, J. G. Characteristics of the question as determinants of dependability. Journal of Consulting Psychology, 1941, 5, 164-169.

c. Blankenship, A. B. Psychological difficulties in measuring consumer preferences. Journal of Marketing, 1942, 6, 66-75.

d. Payne, S. L. The art of asking questions (Rev. ed.). Princeton, N. J.: Princeton University Press, 1963.

1. Formulation of the Question or Question Stem

a. General comments regarding items and question stems. Issues that should be noted concerning the general structure of questions and question stems are:

(1) Question stems may be in the form of an incomplete statement, where the statement is completed by one of the response alternatives, or in the form of a complete question. See Figure VI-C-1 for examples.

Figure VI-C-1

Example of Question Form and
Incomplete Statement Form of Stem

1. How qualified or unqualified for their jobs are most Army NCO's? (Check one.)

_____ Very well qualified

_____ Qualified

_____ Borderline

_____ Unqualified

_____ Very unqualified

2. Check one of the following. Most Army NCO's are:

_____ Very well qualified for their jobs.

_____ Qualified for their jobs.

_____ Borderline.

_____ Unqualified for their jobs.

_____ Very unqualified for their jobs.

The choice between these two methods should depend on which of the two permits simpler and more direct wording for the item in question. Not all of the items in a questionnaire need to be in the same form.

(2) All questionnaire items should be gramatically correct.

(3) All stems should be as neutrally expressed as possible, and the respondent should be permitted to indicate/select the direction of his preference. If this is not done, the stems may influence the response distribution. If items cannot be expressed neutrally, then alternate forms of the questionnaire should be used.

(4) A respondent may not answer an item if he is not able to give the information requested. Therefore, care should be exercised in the wording of the question, so that it does not call for information not possessed by the respondents.

b. Accuracy and completeness of question stems.

(1) The stem of an item should be accurate, even though inaccuracies may not influence the selection of the response alternative.

(2) The question stem, in conjuction with each response alternative, should present the question as fully as necessary to allow the respondent to answer. It should not be necessary for the respondent to infer essential points. An example of an insufficiently informative question stem is given as item 1 in Figure VI-C-2. It is insufficient in that no specification is given as to who should carry the scopes. (The response alternatives are also insufficient since the respondent is not allowed to say "None.") Two or three questions might be needed to obtain all the information desired. Item 2 in Figure VI-C-2 is one revision that makes the question stem sufficient.

(3) Generally, materials which are common to all response alternatives should be contained in the stem, if this can be done without the need for awkward wording.

(4) In forming questions which depend on respondents' memory or recall capabilities, the time period a question covers must be carefully defined. The "when" should be specifically provided.

Figure VI-C-2

An Insufficiently Detailed Question Stem, Plus Revision

1. How many starlight scopes should be issued to a rifle squad?

_____ 1

_____ 2

_____ 3

_____ 4

_____ 5

2. Place a check in front of each squad member's "name" below that you believe should be issued a starlight scope:

_____ Squad Leader

_____ Fire Team 1 Leader

_____ Automatic Rifleman

_____ Grenadier

_____ Rifleman

_____ Fire Team 2 Leader

_____ Automatic Rifleman

_____ Grenadier

_____ Rifleman

_____ Rifleman

(5) Question stems and response alternatives should be worded so that it is clear what the respondent meant. Consider the question "Should this cap be adopted, or its alternate?" If the respondent answers "Yes," it would still be unclear which cap ("this cap" or its alternate) should be adopted.

c. Positive versus negative wording.

(1) Alternative wording can produce demonstrable effects on survey results.

(2) There may be a tendency for the direction of the question stem to be chosen in the response alternative.

(3) Studies have indicated that it is usually undesirable to include negatives in question stems (unless an alternate form with positives is also used for half of the respondents).

(4) Questions worded in positive terms are preferable to questions in negative terms (if alternate forms are not being used). Questions worded negatively may be confusing, or negative words may be overlooked.

(5) If it seems necessary to have a particular question in negative form, the negative word (e.g., not, never) should be underlined or italicized. Care should also be taken that there are no double negatives, as they are frequently misinterpreted.

(6) A question worded in negative terms can often be improved by rephrasing it in positive terms.

d. Definite versus indefinite article wording. The indefinite articles, "a" or "an," would be used in a question such as "Did you see a demonstration of the new night vision device?" A comparable question using the definite article "the" would be, "Did you see the demonstration of the new night vision device?" There is some evidence that changing from "a" to "the" reduces the level of suggestibility of an item. However, there is not enough evidence to warrant a firm conclusion.

e. First, second, and third person wording. An example of a statement written in the first person is, "Army NCO's are understanding of my needs and problems." A statement in the second person is, "Army NCO's are understanding of your needs and problems," while one in the third person is, "Army NCO's are understanding of the needs and problems of their men." It is preferable that the framework of questions be consistent for all questions in a questionnaire, so that responses are comparable. A respondent's opinion of the effects of events affecting his own person is often quite different than his opinions of the effects of the same events on others. Hence, questions written in the first or second person may elicit entirely different responses than the "same" question written in the third person.

There are occasions where each person (first, second, or third) is appropriate. For example, the third person should probably be used when it is desired to elicit information that might be considered too personal for a person to answer about himself. The third person may also be used in attempts to elicit information about the feelings inherent in a minority of respondents, but about which many more respondents may be aware, such as in the statement, "The Army is ahead of most areas of civilian life in reducing racial discrimination." In other cases the first or second person form is not applicable, such as in "The Army is essential for the defense of the country." Also, the use of the third person permits a far larger number of personnel to answer the questions, since some first person questions that are inapplicable to many individuals become applicable when in the third person. Instances may occur where a respondent is asked a question twice, once to discover how he personally feels about the issue (using first or second person), and then to discover what he judges others' feelings on that issue are (using the third person). Generally, however, the use of the third person appears preferable.

f. Loaded and leading questions. Loaded and leading questions should be avoided. Although the questionnaire writer may not deliberately attempt to distort the distribution of responses, he may sometimes do so unintentionally.

In Figure VI-C-3, item 1 should be revised to maintain neutrality by removing the adjectives applied to the rifles. It is true that the M-16 weighs less and fires more rounds faster, but there are other characteristics (accuracy, lethality given a hit, etc.) that are not cited. Hence, the question is loaded because it only presents some of the data relevant to comparing the rifles.

Items 2 and 3 in Figure VI-C-3 show loading of a different type. In item 2, analysis of the available alternatives leaves the impression that the writer of the question thinks at least some should not have a full automatic selector. Analysis of the alternatives in item 3 leads to the suspicion that the writer of the question believes there should be at least one grenade launcher in the rifle squad, since a response alternative of zero grenade launchers was not provided.

There are many additional ways that questions can be loaded. One way is to provide the respondent with a reason for selecting one of the alternatives, as with the question, "Should we increase taxes in order to get better schools, or should we keep them about the same?" A question can also be loaded by referring to some prestigious individual or group, as in, "A group of experts has suggested...Do you approve of this, or do you disapprove?"

Figure VI-C-3

Examples of Loaded Questions

1. Which rifle do you prefer, the lighter, faster shooting M16 or the heavier, slower firing M14?

    _____ M16

    _____ M14

2. Should every rifleman in the rifle squad have a full automatic selector on his rifle?

    Yes_____

    No______

    If no, how many should?_____

3. How many grenade launchers (M79) do you desire in the rifle squad?

    _____ 1

    _____ 2

    _____ 3

    _____ 4 or more

Leading questions are similar to loaded questions. Two examples are shown in Figure VI-C-4. The problem is that most people are reasonably cooperative and like to help. If they can figure out what is wanted, they will often try to comply. The items in Figure VI-C-4 were actually used in the collection of data in a field test. As might be expected, the impression received from an analysis of the results is that men are, in general, highly motivated, and use good noise discipline during movement. (These items also allow respondents to avoid critizing, and to give socially desirable answers.)

Figure VI-C-4

Examples of Leading Questions

1. Do you think your men were pretty highly motivated on this exercise?

   Yes_____

   No_____

2. Were they pretty good at using good noise discipline during movement?

   Yes_____

   No_____

The best way to avoid loaded questions is to find a devil's advocate to review them or to pretest the items on someone who holds opposite or minority views. Another check is to ask yourself what you think, what someone who disagrees with you would think, and whether your response alternatives would give him a chance to present his views.

There are times when loaded questions probably should be used. This is when, without loading, the question would pose an ego-threat to the respondent, so that he might give an untruthful reply. The loading removes the ego-threat so that a more valid response can be obtained. An example might be, "Many people are not able to get as much schooling as they would like. What was the last grade you completed in school?"

g. Embarrassing or self-incriminating questions. Respondents should not be asked embarrassing or self-incriminating questions. Consider the question, "Did you clean your weapon regularly in Vietnam?" It is asking respondents who did not clean their rifles regularly to expose themselves to possible embarrassment. Thus, one would expect the percentage of "No" responses to fall short of the true percentage not cleaning their weapons "regularly."

h. Questions that ask respondents to go against basic inclinations.

Many people are reluctant to criticize, though they enjoy giving praise. Thus, a question that allows a respondent to avoid criticism will bias his answers; similarly, a question that offers him the opportunity to criticize may bias responses because he will not wish to do so. Figure VI-C-5 illustrates this.

Figure VI-C-5

Example of a Question
Asking the Respondent to Criticize

1. Was your unit's use of fire and maneuver correct, and in accordance with current Army doctrine?

Yes_____

No______

If no, why not?____

____

The question in Figure VI-C-5 asks the respondent either to criticize his unit or to avoid criticism. Some respondents might answer "No," if they have an important point to make. However, a substantial number of others will wash their hands of the whole affair and answer "Yes," although they might feel that performance was not completely correct.

i. Inclusion of different subjects into the same question. Double-barreled (compound) questions, in which a respondent can agree with one part of a question and disagree with another, should be avoided. Consider, for example, item 1 in Figure VI-C-6. Most respondents would probably want to rate completeness and accuracy differently, since in most situations research has shown that they are negatively correlated. Therefore, ratings of the two aspects of performance should be rated separately, as shown in items 2 and 3 of Figure VI-C-6.

Figure VI-C-6

Examples of Double-Barreled Questions and Alternatives

1. How complete and accurate was the surveillance information?

_____ Very satisfactory

_____ Satisfactory

_____ Borderline

_____ Unsatisfactory

_____ Very unsatisfactory

2. How complete or incomplete was the surveillance information?

_____ Very complete

_____ Fairly complete

_____ Borderline

_____ Fairly incomplete

_____ Very incomplete

3. How accurate or inaccurate was the surveillance information?

_____ Very accurate

_____ Fairly accurate

_____ Borderline

_____ Fairly inaccurate

_____ Very inaccurate

It may be noted that in item 2 of Figure VI-C-6 both "complete" and "incomplete" are included. Similarly, both "accurate" and "inaccurate" are in the stem of item 3. To use only one (e.g., "complete") in the stem would tend to inflate the number of respondents selecting that alternative.

j. Use of giveaway words. Avoid words which lead the careful thinker to respond in the negative while others, thinking less carefully, respond in the positive. Consider for example the question, "Do you feel that your unit did its best in all contacts over the past six months?" One wonders if any unit can do its actual best, except very rarely. The word "all" makes this an even more difficult question to answer positively.

k. Ambiguous questions. Vague or ambiguous words or questions should be avoided. For example, the question "What is your income?" is not sufficiently specific. The respondent may give monthly or annual income, income before or after taxes, his income or the family income, etc.

As another example, consider item 1 in Figure VI-C-7.

Figure VI-C-7

Example of Ambiguous Question and Alternative

1. Did you clean your rifle regularly in Vietnam?

_____ Yes

_____ No

2. How often, on the average, did you clean your rifle in Vietnam?

_____ Every day

_____ Once every two days

_____ Once every three days

_____ Once every four days

_____ Other (please specify): ______________________________

____

Use of the word, "regularly" without specification of the time interval between cleanings is a defect in the question. A respondent could justify a "yes" by thinking to himself: "Sure, I cleaned it regularly - once every four months."! Because of the self-exposure involved, the questionnaire item approach to this topic is probably not capable of providing an accurate estimate, but rewording could still make the amount of underestimation less. So, if the data cannot be collected by field inspection, the revised questionnaire item could read like item 2 in Figure VI-C-7.

2. Formulation of the Response Alternatives

When formulating the response alternatives portion of a questionnaire item, the following points should be kept in mind:

a. All response alternatives should follow the stem both gramatically and logically, and if possible, be parallel in structure.

b. If it is not known whether or not all respondents have the background or experience necessary to answer an item, (or if it is known that some do not), a "Don't know" response alternative should be included.

c. When preference questions are being asked (such as (Which do you prefer, the M16 or the M14 rifle?") the "No preference" response alternative should usually be included. The identification of "No preference" responses permits computation of whether or not an actual majority of the total sampled are pro or con.

d. The use of the "None of the above" option or variants of it such as "Not enough information" is sometimes useful.

e. The option "All of the above" may on rare occasions be useful. It seems more appropriate to academic test questions than to the questioning of field test participants.

f. For most items, the questionnaire writer desires the respondent to check only one response alternative. Use of the parenthetic "(Check one.)" should eliminate the selection of more than one alternative. It is very important to make it clear to the respondent that he may check more than one alternative in those fairly rare instances where the questionnaire writer does wish to permit this.

g. In some instances, response categories as long as a sentence may be more desirable than short descriptors. In rare cases, numbers may be used without verbal descriptors, if the numbers have been previously defined.

h. Number of response alternatives is discussed in Section VI-G, order of response alternatives in Section VI-H, response anchoring in Chapter VII, and the order of perceived favorableness of commonly used words and phrases in Chapter VIII.

3. Expressing Directionality and Intensity in Stem Versus Response Alternatives

In item 1 of Figure VI-C-8, directionality (in this case, satisfaction) is expressed in the question stem.

Figure VI-C-8

Alternate Ways of Expressing Directionality and Intensity

1. The M16 is a satisfactory rifle.

_____ Agree

_____ Disagree

2. The M16 is

_____ a satisfactory rifle.

_____ an unsatisfactory rifle.

3. The behavior of civilian employees of the PX toward enlisted personnel is extremely offensive.

_____ Agree

_____ Disagree

4. The behavior of civilian employees of the PX toward enlisted personnel is

_____ very offensive.

_____ somewhat offensive.

_____ neutral.

_____ somewhat pleasant.

_____ very pleasant.

In item 2 the directionality is expressed in the response alternatives. In item 3 the stem contains terms of intensity and directionality, while these terms are located in the response alternatives in item 4. Item 2 is preferred to item 1, and item 4 is strongly preferred to the item 3 approach.

The rationale for this preference is similar to the discussion of positive versus negative terms. Those who check "Disagree" to item 3 have not been permitted to indicate what it is they would agree with, (e.g., those who feel employees are offensive but not extremely offensive would have to check "Disagree" as would those who feel employees are very pleasant), whereas the construction of item 4 does permit them to do so. It would take five versions of item 3 to correct this deficiency and achieve the coverage of opinion incorporated by the response alternatives of item 4.

D. Difficulty of Items

1. One of the major recommendations advanced by almost every general source on how to write sound questionnaires is "keep it simple." Logic dictates that words used in surveys should not have multiple meaning, nor should they be beyond the level of vocabulary of the typical respondent. Words, phrases, and sentence structures that the respondent can understand should be used.

Consider item 1 in Figure VI-D-1. It contains too many hard to understand words. Many respondents would have difficulty understanding either the question or the response alternatives. In the revision in item 2, the words have been simplified, and a "catch-all" open-ended response alternative added (to catch all other reasons).

Figure VI-D-1

Example of Hard to Understand Item and Alternative

1. In the highly specialized counterinsurgency environment represented by the basically internecine affair in Vietnam, what would you say should represent the basic essence of our rationale for continuation of our involvement?

_____ Prolongation of attrition of enemy forces, in order to reduce the level of threat to South Vietnam.

_____ Orderly transfer of military responsibility to the host country, in order to produce stabilized competency to deal with any future internal disturbances.

2. What is our main reason for staying in Vietnam? (Check one)

_____ To reduce the threat to South Vietnam by continuing the destruction of enemy forces.

_____ To assure South Vietnam's survival while it takes over responsibility for its own protection.

_____ Other (specify) ____________________________________

____

It should not be assumed that the respondent will understand what the question writer is talking about. Consider, for example, the question "Which do you prefer, dichotomous or open questions? The odds are that a fairly substantial number of people would not be able to define these two question types. However, if they are asked this question, they will be happy to choose. The point is that people will not volunteer their ignorance of something, though they may admit it if you ask them. However, this caution goes beyond ignorance of an issue. Another problem is that the specialist wording the question may simply have an unusual command of his own language. Scientific jargon has been criticized. Perhaps overlooked is the fact that there are other kinds of jargon, too. The question asker has a responsibility to make himself understood. One way of screening for individuals who do not have a basis for providing the information needed is to include one or two pure information questions, planning to discard questionnaire returns from respondents who cannot answer the information questions correctly. However, our usual policy should be to throw out or revise items that are not understandable, rather than to throw out the responses of the people who can't understand the item.

2. Ways of Measuring Item Difficulty

Various procedures exist for determining the difficulty or reading comprehesion level of printed material. Such a discussion is, however, beyond the scope of the preliminary version of this manual. Sources that may be consulted include:

a. Dale, E., & Chall, J. S. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-20, 37-54.

b. Flesch, R. A new readability yardstick. Journal of Applied Psychology, 1948, 32, 221-233.

c. Fry, E. A readability formula that saves time. Journal of Reading, 1968, 11, 513-516.

d. Lorge, I. Predicting readability. Teachers College Record, 1944, 45, 404-419.

e. Thorndike, E. L., & Lorge, R. The teacher's word book of 30,000 words. New York: Columbia University Press, 1944.

E. Length of Question/Stem

This section notes some considerations about the length of question stems. There is little research in this area to guide the questionnaire writer. See Section IX-C regarding questionnaire length.

1. It is sometimes desirable to break the question stem into two or more sentences when the sentence structure would otherwise be unnecessarily complex. For instance, one sentence can state the situation, and one can pose the question. Lengthy question stems that try to explain a complicated situation to the respondent should be avoided. If the respondent is not aware of the facts presented, he may become more confused or biased than enlightened, and his opinion would not mean much.

2. Longer open-ended questions do not necessarily produce a greater amount of and more accurate information than shorter ones. However, it may take more words to achieve a proper focus.

3. Questionnaire developers have a tendency to use long question stems with true-false questions when "True" is the correct answer. Respondents often detect and react to this tendency. Field test questionnaires, however, should make relatively little use of "True" and "False" response alternatives. These alternatives are more appropriately used when testing whether respondents have acquired a required proficiency level, for example, the ability to visually recognize a given type of enemy aircraft.

F. Order of Question Stems

There are two issues to consider regarding the order of question stems. The first has to do with the order of questions within a series of items that are designed to explore the same topic or subject matter or related subject matter areas. The second has to do with the order of different groups of questions when the groups deal with fairly separate topics or subject matter areas. For example, one group of questions may deal with factual items, while another may deal with attitudes. If items bearing on the same point are presented in succession, the respondent can proceed more readily through them. Thus this is usually a desirable practice. An exception arises when one wishes to check the consistency of the respondent. To do this, two (or more) similar items are included, but at widely different points in the questionnaire.

1. Order of Questions Within a Series of Items

a. It is often recommended that the order of questions on a instrument be varied or assigned randomly to avoid one question contaminating another. The view is that the immediately preceding question or group of questions places the respondent in a "mental set" or frame of reference. For example, asking respondents a general question about their feelings regarding automobile exhaust pollution might influence responses to the question, "Do you prefer leaded or nonleaded gasoline?" Although this effect may be prominent in specific settings or with specific questionnaires, there is little evidence in the literature to support its general existence.

b. Sometimes it is recommended that broad questions be asked before specific questions. The rationale for this approach is that the respondent can more easily and validly answer specific questions after having had a chance to consider the broader context. Also, asking the specific questions first could influence the response to the broader question. Sometimes, however, it is best to start with the more specific questions, especially when the respondent should have experiences or issues in mind when he answers the more general questions; or when the questionnaire deals with a complex issue which the respondent may not have thought too much about.

c. The order of questions within a series of items will also depend upon whether filter questions are needed. A filter question is used to exclude a respondent from a particular sequence of questions if those questions are irrelevant to him. For example, if a series of items were asked about different kinds of weapons, a "No" response to a question such as "Have you ever used the M14 rifle?" might be used to indicate that the respondent should skip the following question(s) about the M14.

2. Order of Different Groups of Questions

   a. There is usually a psychological or logical order in which to ask the questions, so that the questionnaire flows smoothly from one topic to the next and the respondent is not shifted frequently from one topic to another and back again. However, a shift from one topic to another should be apparent to the respondent.

   b. It is usually recommended that more difficult or more sensitive questions be asked later in the questionnaire, possibly at the end.

   c. One or more easy, nonthreatening questions should probably be asked first to build rapport. They should be short and easy to understand and to answer. But they should not be irrelevant to the objectives of the questionnaire. Verbal efforts to build rapport by the questionnaire administrator seem preferable to using questionnaire content.

3. Effects of Order of Questions on Subjects' Responses

There is no evidence that the order of presentation of questions on a questionnaire has any effect on the subject's choice of response alternatives.

G. Number of Response Alternatives

One of the basic issues in the use of rating questions or attitude scales is the determination of the optimum number of responses, alternatives or categories. Researcher's habit or tradition rather than solid empirical support often has led to the recurrent use of five-point rating scales, seven-point semantic differential scales, and so on. The reason for concern with the number of response alternatives stems from the belief that a "coarse" scale with too few response alternatives may result in a loss if information concerning the respondents' discrimination powers. It may reduce the respondents' cooperation in rating, as a coarse scale "forces" judgments and thereby irritates some respondents. An extremely "fine" scale, with too many response alternatives, may go beyond the respondents' powers of discrimination, be excessively time consuming, or difficult to score.

The following sections consider number of response alternatives to use in multiple choice, rating scale, and forced choice items: Section VI-C-3 - formulation of response alternatives; Section VI-H - order of response alternatives; Chapter VII - response anchoring; Chapter VIII - order of perceived favorableness of words and phrases.

1. Number of Response Alternatives with Multiple Choice Items

No firm rules can be established regarding the number of response alternatives to use with multiple choice items. It depends in a large part upon the question being asked and the number of answers logically possible. The following considerations, however, may be noted:

a. There is some evidence that dichotomous items (items with only two response alternatives) are statistically inferior to items with more than two response alternatives.

b. Dichotomous items are easier to score than nondichotomous items, but they may not be accepted as well by the respondent.

c. A good nondichotomous multiple choice item usuallly can not be written as a set of separate dichotomous items.

d. Consideration should be given to the fact that many response alternatives may make a questionnaire unduly time consuming.

e. The number of choices logically possible or desirable should constitute an upper limit on the number of response alternatives used for an item.

f. Non-existent response alternatives may be checked by the respondent if an answer sheet is used which has more spaces than there are alternative answers, e.g., the answer sheet has five spaces for each question but some questions have fewer than five alternatives.

2. Number of Response Alternatives with Rating Scale Items

Authorities in psychometrics contend that the optimal number of response alternatives to employ with rating scales is a matter for empirical determination in any situation. They also suggest that considerable variation in number around the optimal number changes reliability very little. These conclusions seem to be supported by the available research literature. Although rules regarding the number of response alternatives to use with rating scales cannot, therefore, be firmly established, the following issues can be considered.

a. The effects of increasing or decreasing the number of response alternatives for a question cannot be generally specified with certainty. Increasing the number of response alternatives does not necessarily increase reliability, and there is no consistent relationship between the number of response alternatives and validity.

b. J. P. Guilford (in Psychometric methods. New York: McGraw-Hill, 1954) reported that seven response alternatives is usually lower than optimal, and it may pay in some favorable situations to use up to 25 scale divisions. Others believe that seven steps or five is optimal. Some believe that five should be used for single or unipolar (one direction) scales, nine for double or bipolar scales. Many practitioners consistently use five-point scales. Sometimes a nine-point hedonic (pleasure) scale is recommended for food items, and a six-point scale for other uses.

c. The number of response alternatives to use is often determined on the basis of the degree of discrimination required. For example, a nine-point scale may sometimes (but not always) give greater discrimination than a three-point scale.

d. Psychologists with considerable experiences in military operational field testing feel that anything more than five alternatives is too great a number for many junior enlisted personnel to discriminate among. More non-responses are secured and the reliability of discrimination of answered items is not increased.

e. Questionnaire administration time is probably a function of the number of response alternatives.

f. There is some evidence that increasing the number of response alternatives seems to decrease the number of nonresponses and uncertain responses (e.g., "Cannot decide").

g. In addition to the response alternatives representing the rating scale continuum, it may be necessary to add alternatives such as "Have no effect" or "No opinion."

h. Scoring and data analysis considerations may affect the selection of the number of response alternatives. If Chi square tests are sufficient, two or three response alternatives might be adequate. However, if two or three response alternatives are used when nonparametric rank order correlations are employed, substantial "ties" on ranks will result. If parametric statistics are to be employed, more alternatives are usually better, because of the assumption of continuous distributions or interval scale properties.

3. Number of Response Alternatives with Forced Choice Items

A number of different forced choice item formats have been used, such as the following:

a. Two phrases or statements per item, both favorable or both unfavorable, choose the more descriptive or the least descriptive.

b. Three statements per item, all favorable or unfavorable, choose the most and least descriptive statements in each item.

c. Four statements per item, all favorable, choose the two most descriptive statements.

d. Four statements per item, all favorable, choose the most and least descriptive statements.

e. Four statements per item, two favorable and two unfavorable, choose the most and least descriptive statements.

f. Five statements per item, two of which were favorable, one neutral, and two unfavorable in appearance, choose the most and least descriptive.

The evidence is not clear, but three or four statements per item may be preferable to two. One study concluded that the format described in "c" above was superior to the others. It was most bias resistant, yielded consistently high validities under various conditions, had adequate reliability, and was one of the best recieved by respondents.

H. Order of Response Alternatives

1. General Considerations

The experimental evidence on the effect that the order of presentation of response alternatives for a question has on a subject's choice of response is inconclusive and contradictory. Varying conclusions include:

a. Respondents have a tendency to select the first response alternative in a set more than the others.

b. With multiple choice questions there is tendency to choose answers from the middle of the list, if the list consists of numbers, and from either the top or bottom of the list, if the alternatives are fairly lengthy expressions of ideas.

c. Poorly motivated respondents tend to select the center or neutral alternatives with rating scale items.

d. On items about which respondents feel strongly the order of alternatives makes no difference. On items about which the respondent does not feel strongly, most will tend to check the first alternative.

e. The positive pole of rating scale response alternatives should be presented first since this will improve the reliability of the responses. However, it is important to realize that reliability may increase while validity decreases.

Test item form biases are discussed in Section XII-B.

2. Suggested Order for Multiple Choice Items

The following suggestions are offered regarding the order of multiple choice items:

a. When the response alternatives have an immediate apparent logical order (e.g., they all relate to time) they should be put in that order.

b. When the response alternatives are numerical values, they should in general be put in either ascending or decreasing order.

c. When the response alternatives have no immediately apparent logical order, they should generally be put in random order.

d. Alternatives such as "None of the above" or "All of the above" should always be in the last position.

e. Alternate questionnaire forms (e.g., where the order of alternatives is reversed on half of the forms) are often desirable.

3. Suggested Order of Rating Scale Items

Since rating scales call for the assignment of objects along an assumed continuum or in ordered categories along the continuum, it follows that the response alternatives must be in order from "high" to "low" or "low" to "high", with the choice of words for "high" and "low" (the end point labels) depending upon the continuum being used. For example, for the continuum satisfactory-unsatisfactory, item 1 in Figure VI-H-1 uses the "high" to "low" order, while item 2 uses the order "low" to "high".

Figure VI-H-1

Example of Rating Scale Item
with Alternate Response Alternatives Order

1. The M16 rifle is:

_____ very satisfactory.

_____ satisfactory.

_____ borderline.

_____ unsatisfactory.

_____ very unsatisfactory.

2. The M16 rifle is:

_____ very unsatisfactory.

_____ unsatisfactory.

_____ borderline.

_____ satisfactory.

_____ very satisfactory.

Many practitioners use the "high" to "low" order. If one has reason to believe that the order of the response alternatives makes a difference, or wishes to make certain that they do not, then the use of alternate questionnaire forms is recommended. Each alternate form should list the response alternatives in a different order. The "good" or "high" end of the scales should be at the same end of each scale for all items in a given questionaire form, but the order should normally be reversed on 50% of the forms. For example, the order shown in item 1 in Figure VI-H-1 would be used on half of the forms, the order shown in item 2 on the other half. (Normally, there would be only two questionnaire forms, one with each order, but at times alternate forms are also needed for other purposes. Hence, there may be more than two.)

Chapter VII: Response Anchoring

A. Overview

This chapter has to do with the "anchoring" of rating scale responses, that is, with the words used to define some or all of the response alternatives. Section VII-B shows various types of response anchors, while Section VII-C discusses anchored versus unanchored scales. The amount of verbal anchoring is the topic of Section VII-D, while some procedures for the selection of verbal scale anchors are presented in Section VII-E. Finally, Section VII-F discusses balanced versus unbalanced scales.

It should be noted that Section VI-C 3 discussed the formuation of response alternatives, while the number and order of response alternatives are the topics of Sections VI-C and VI-H, respectively. The order of perceived favorableness of words and phrases is discussed in Chapter VIII.

B. Types of Response Anchors

There are a number of different types of response anchors that can be used with rating scale items. Some have been shown as examples in other chapters, such as Section VI-D. Five other types of response anchors are shown in Figure VII-B-1. The first shows the original form of the semantic differential. It is a combination graphic and verbal scale. Respondents were instructed to place an "X" on the line that represented their attitude. The use of verbal anchors with a -5 through +5 numerical continuum is shown in item 2 of Figure VII-B-1. Item 3 shows verbal anchors used with a 1 through 11 numerical continuum. A combination verbal and numerical continuum is shown in item 4, while a verbal continuum is shown in items 5 and 6. Item 6 is a typical Likert rating scale that calls for a verbal rating to a directional statement that may be phrased either positively or negatively. An example might be "The Modern Volunteer Army places too much emphasis on extrinsic factors (such as beer in the barracks) as opposed to intrinsic, job related factors (such as pay or supervision)."

Sufficient empirical support exists to conclude that the reliability of scales with verbal anchors and verbal response alternatives is superior to that of purely numerical scales.

Figure VII-B-1

Types of Response Anchors

1. Combination graphic and verbal scale.

Strong_____: _____: _____: _____ :_____: _____: _____ :_____: _____: Weak

2. Verbal anchors with a -5 through +5 numerical continuum.

Definitely dislike   Definietely like

-5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

3 Verbal anchors with a 1 through 11 numerical continuum.

Definitely dislike   Definitely like

1 2 3 4 5 6 7 8 9 10 11

4. A verbal and numerical contiuum.

Dislike completely	Dislike somewhat	Dislike a little	Neither like nor dislike	Like a little	Like somewhat	Like completely
1	2	3	4	5	6	7

5. A verbal continuum.

Below average	About average	A little better	A lot better	One of the best	None better

6. A verbal continuum. (Likert rating scale)

__Agree strongly __Agree __Undecided __Disagree __Disagree strongly

C. Anchored Versus Unanchored Scales

A number of studies have been conducted on the topic known as "anchoring effects." Unfortunately, the research evidence is contradictory as to whether anchored or unanchored scales should be used. It has been noted that unanchored scales may well be anchored by the question stem, so that the response alternatives may not have to be. When only one end of a scale is anchored, some studies have found a tendency for respondents to move toward that extreme. But other studies have found the opposite tendency. At least one study found that judgment time is decreased with anchoring. In practice, then, it is usually best to use anchored scales.

D. Amount of Verbal Anchoring

Obviously the amount of verbal anchoring of a rating scale item can vary. It can be anchored at the center, or on the ends or both, or at many points on the entire continuum. There is some evidence that more descriptive data can be obtained with more anchoring, and that greater scale reliability is achieved with added verbal anchoring. Scales with verbal descritpors for all response alternatives may also be better predictors of behavior. On the other hand, adding examples to definitions does not seem to help too much. (See also Section VI-G regarding the number of response alternatives to employ.)

E. Procedures for the Selection of Verbal Scale Anchors

Some guidance can be offered regarding the selection of verbal scale anchors. See also Chapter VIII.

1. Scales can be anchored by examples of expected behavior based upon observations of behavior.

2. Pretests for the selection of verbal anchors are valuable in building scale content. Rather than employing anchors which seem appropriate, anchors should preferably be selected by respondents similar to those who will be participating in the study.

3. Scale endpoints that are unrealistically extreme, such that few if any respondents would select them, should be avoided. For example, it may be seldom that "Never" or "Always" apply, so that the use of "Rarely" and "Usually" may be more appropriate. There are instances however, where extreme statements are realistic. The decision here often requires experience with what is being rated.

4. Analysis of data is normally facilitated if verbal scale anchors selected for rating scales are of equal distance from each other in terms of scale values. See, however, Chapter, VIII.

F. Scale Balance, Midpoints, and Polarity

1. Balanced Versus Unbalanced Scales

Historically, balanced scales have been preferred by researchers. A scale is balanced when it has a number of positive response alternatives equal to the number of negative alternatives, regardless of the presence or absence of an "indifferent" or neutral category. A "Don't know" response alternative, if present, is not considered to be part of the scale, so is not counted when deciding if the scale is balanced. See the examples of balanced and unbalanced scales in Figure VII-F-1. Unbalanced scales may be employed if pretest results indicate that many respondents will be choosing extreme response alternatives at one end of a scale, producing a skewed distribution of responses rather than the statistically expected normal distribution around the mean attitude. To reduce the piling up of responses at one end of a scale, - or, to add to your ability to discriminate among responses in that region - the scale is made unbalanced by adding more response alternatives on the side of the scale where the piling is likely to occur. This practice tends to spread the distribution of responses more evenly along the scale continuum.

In cases where one has no advance information or other basis for expecting responses to be largely one-sided, it is normally desirable to have an equal number of positive and negative response alternatives; i.e., a balanced scale.

2. Midpoints

Scales may or may not include a midpoint or neutral response alternative; this does not affect their classification, but does affect their response distributions. As examples, items 1c, 2a, and 3 in Figure VII-F-1 show scales with no neutral point. One might exclude the neutral point for items where it is judged that respondents ought to have a sufficient basis for being pro or con and where one desires to force respondents away from an "on the fence" position. Bipolar scales should be balanced in terms of the degree of extremeness denoted by the end point anchors. For example, if "Never" is used, then "Always" should be used as the opposite end point.

3. Polarity

Scales may be bipolar or unipolar. Item 3 in Figure VII-F-1 illustrates a unipolar scale. Its basic feature is that it represents the thing being assessed as having from none to a

Figure VII-F-1

Examples of Scale Balance, Midpoints, and Polarity

1. Balanced bipolar scales.

   a. Very progressive
      Progressive
      Moderately progressive
      Neither progressive nor conservative
      Conservative
      Very conservative

   b. Effective
      Fairly effective
      Borderline
      Fairly ineffective
      Ineffective

   c. Very effective
      Somewhat effective
      Somewhat ineffective
      Very ineffective

   d. Very satisfied
      Satisfied
      Borderline
      Dissatisfied
      Very dissatisfied

2. Unbalanced bipolar scales.

   a. Enthusiastic
      Extremely favorable
      Very favorable
      Favorable
      Fair
      Poor

   b. Quite good
      Rather good
      Somewhat poor
      Rather poor
      Quite poor
      Very poor

3. Unbalanced Scale (unipolar).

   Very much
   Much
   Some
   A little
   None

maximum - with n steps in between - of some property. The question of balance only arises for bipolar scales. Many a bipolar scale could be re-designed as a unipolar scale. Instead of item 1c in Figure VII-F-1, one's question about effectiveness (not given) could have been followed by this unipolar scale of effectiveness: maximum effectiveness, great effectiveness, moderate effectiveness, slight effectiveness, and no effectiveness.

Semantic preferences may determine whether the questionnaire writer uses bipolar or unipolar scales.

Chapter VIII: Empirical Bases for Selecting Modifiers for Response Alternatives

A. Overview

When constructing a questionnaire, it is often necessary to select adjectives, adverbs, or adjective phrases to use as response alternatives. The words selected for response alternatives should be clearly understood by the respondents to the questionnaire and they should have precise meaning. There should be no confusion among respondents as to whether one term denotes a higher degree of favorableness or unfavorableness than another.

There is no need to guess which phrases or words are the best to use as response alternatives. Many studies have been conducted in order to determine the perceived favorableness of commonly used words and phrases. These studies have determined scale values and variances for words and phrases which can be used to order the responsive alternatives. In some of the studies ambiguous words and words that are not appropriate to use as response alternatives have been identified.

The results of these studies and the experience of questionnaire designers have been incorporated into this chapter in order to offer guidelines and suggestions to be used in selecting response alternatives. This chapter includes lists of words and procedures to use in selecting response alternatives. Many lists of phrases with mean scale values and standard deviations are presented. The scale values are given for the purpose of selecting response alternatives, not for the purpose of assigning scale values to response alternatives for data analysis purposes.

Section VIII-B discusses things to consider in selecting response alternatives; Section VIII-C covers the selection of response alternatives denoting degrees of frequency; Section VIII-D, the selection of response alternatives using order of merit lists of descriptor terms; Section VIII-E, the selection of response alternatives using scale values and standard deviations. Section VIII-F includes sample sets of response alternatives.

Scale values, standard deviations, and interquantile ranges reported in this chapter have been taken from data presented in the following studies:

1. Altemeyer, R. A. Adverbs and intervals: A study of Likert scales. Proceedings of the Annual Convention of the American Psychological Association, 1970, 5(pt. 1), 397-398.

2. Cliff, N. Adverbs as multipliers. Psychological Review, 1959, 66, 27-44.

3. Dodd, S. C., & Gerberick, T. R. Word scales for degrees of opinion. Language and Speech, 1960, 3, 18-31.

4. Gividen, G. M. Order of merit- descriptive phrases for questionnaires. Fort Hood Texas: OCRD Army Research Institute Field Unit, 22 February 1973.

5. Jones, L. V., & Thurstone, L. L. The psychophysics of semantics: An experimental investigation. Journal of Applies Psychology, 1955, 39, 31-36.

6. Matthews, J. J., Wright, C. E., & Yudowitch, K. Analysis of the results of the administration of three sets of descriptive phrases. Palo Alto: Operations Research Associates, March 1975.

7. Mosier, C. I. A psychometric study of meaning. Journal of Social Psychology, 1941, 13, 123-140.

8. Myers, J. H., & Warner, W. G. Semantic properties of selected evaluation adjectives. Journal of Marketing Research, 1968, 5, 409-412.

9. U.S. Army Test and Evaluation Command. Development of a guide and checklist for human factors evaluation of Army equipment and systems. U.S. Army Test and Evaluation Command (TECOM), 1973.

B. General Considerations in the Selection of Response Alternatives

There are several ways of selecting response alternatives. These ways are dependent on the purpose of the questionnaires and/or on the way the data will be analyzed. There are specific considerations when selecting response alternatives for balanced scales, when selecting response alternatives with extreme values, and when developing equal interval scales. There are also general things to consider in the selection of any response alternative.

In some cases it is desirable to select response alternatives on more than one basis. For example, mutually exclusive phrases may be selected also on the bases of parallel wording.

1. Matching the Question Stem

Descriptors should be selected to follow the question stem. For example, if the stem asks for degrees of usefulness, descriptors such as "Very useful" and "Of significant use" should be used. In some cases this may mean rewording the question stem so that appropriate response alternatives can be selected.

2. Mixing Descriptors

Descriptors on different continuums should usually not be mixed. For example, "Average" should never be used with quantitative terms or qualitative terms such as "Excellent" or "Good" (since "average" performance for a group may very well be excellent or good or even poor). If the descriptors are selected for use with a question stem asking about satisfactory or unsatisfactory, the word "Satisfactory" or "Unsatisfactory" (or a synonym) should normally be in every response alternative, except perhaps for a neutral response alternative.

Some experts go as far as to say that the wording of the response alternatives should be parallel for balanced scales. For example, if the phrase "Strongly agree" is used then the phrase "Strongly disagree" should also be used. By reviewing some of the studies that have determined scale values for descriptors, it can be seen that some pairs of parellel phrases are not equally distant from a neutral point or from other phrases in terms of their scale values. Hence, parallel wording may not always provide equally distant pro and con response alternatives, although they may be perceived as symmetrical opposites.

Using descriptors from one continuum or descriptors with parallel wording for a given questionnaire item has advantages. The advantages are that the response alternatives will usually fit the stem better, and they will be parallel to each other in meaning and appearance.

3. Selecting Response Alternatives with Clear Meaning

Some words are difficult for respondents to use in answering questions. This difficulty may be the result of the respondent being ignorant of the meaning of the word, or not being able to rate the word in terms of degrees on specific scales. Such words should not be used as response alternatives. Some studies asked the respondent to indicate which words he was unable to rate. Table VIII-B-1 lists examples of words that were unrateable by subjects.

Table VIII-B-1

Words Considered Unrateable by Subjects

Phrase	Phrase
Adverse	Noxious
Appalling	Peerless
Base	Satiating
Despicable	Seemly
Expedient	Superlative
Fit	

From: Mosier 1941a.

Some words appear to have two or more distinct meanings. When these words are rated on a continuum of favorableness-unfavorableness, many respondents will check around one part of the scale while the other respondents will check around a different place on the scale. It is said that these words produce bimodality of response. Such words also should not be used as response alternatives. A list of words exhibiting bimodality of response is given in Table VIII-B-2.

Table VIII-B-2

Words Exhibiting Bimodality of Response

Phrase	Phrase
Acceptable	Irresistable
Amazing	Normal
Bearable	Tempting
Completely indifferent	Unfit
Extremely indifferent	Unspeakable
Highly indifferent	Unusually indifferent
Important	Very indifferent
Indifferent	Very, very indifferent
Indispensable	

From: Mosier 1941a

4. Selecting Nonambigious Terms

Some descriptors are more ambiguous than others. The more ambiguous the descriptor, the more varied the respondents' interpretations of the degree of favorableness denoted by the descriptor. The ambiguousness of a descriptor is measured by the variability of responses given to the item. One measure of variability is the standard deviation. When available, standard deviations (SD) are given with scale values in this chapter. Another measure used to show variability is the interquartile range. This measure is indicated in this chapter with scale values only when the standard deviations were unavailable.

It is most desirable to select terms with small ranges or small standard deviations, as they will have less ambiguous meaning to respondents. Also, selecting a term with a small standard deviation decreases the chances of the meaning of the term overlapping with the meaning of neighboring terms.

5. Selecting Response Alternatives

When balanced scales with two, three, four, or five descriptors are sufficient for describing the distribution of respondents' attitudes or evaluations, the questionnaire writer can compose them quite satisfactorily by using a term and its literal opposite (effective vs. ineffective; pleasing vs. unpleasing) for two of the terms. A more extreme pair can be produced by using "Very" to modify these two terms.

The first of several intended studies of how people rate/order terms that might be used for rating scale descriptors was conducted by Operations Research Associates and ARI just prior to the writing of this manual. Its results may assist questionnaire developers who need unbalanced scales or scales with more than five descriptors. In the study each of 100 Army personnel was asked to assign a scale value ranging from -5 (most negative) to +5 (most positive) to each term in three different sets of terms, totaling over 100 descriptors.

Tables VIII-B-3 and VIII-B-4 give samples of descriptors from this study for which mean scale values and standard deviations have been calculated. The list in Table VIII-B-3 was derived by first selecting the descriptor with the largest positive mean. The next descritpor selected has a mean that is at least one standard deviation lower. The implication of the gap of one standard deviation is that not more than 16% of the people would have assigned a lower scale value to the first descriptor than they did to the second descriptor, and vice versa. To this extent the raters disagreed on the ordering of these two terms when rating about 50. The third descriptor on the list has a mean scale value yet another standard deviation lower. This process was repeated until the descriptor with the lowest mean scale value was selected. A descriptor was not used if its standard deviation was greater than 1.000.

The list on Table VIII-B-4 was constructed again by skipping at least one standard deviation between adjacent terms; however, the starting point was at the middle, with the word "neutral."

Use of Table VIII-B-3 as a 10-descriptor unbalanced scale is not highly recommended. If one wanted a nine-descriptor scale, he could use the four adverbs appearing in front of "Acceptable" in the table in that same location, and also use them in front of "Unacceptable" in reverse order to create a semantically balanced and ordered scale. Or, one could use the five adverbs, now shown below "Neutral," both above and below "Neutral" to create an 11-descriptor scale of acceptability (or effectiveness, or satisfactoriness, etc.). "Neutral," however, may not be a suitable midpoint term here as the respondent who has neutral feelings (i.e., does not know or does not care) might check this response, whereas the term "neutral" is intended to specify, for example, a midpoint between "barely acceptable" and "barely unacceptable."

Table VIII-B-3

Sample List of Phrases Denoting Degrees of Acceptability

Phrases	Mean	SD
Wholly acceptable	4.725	.563
Highly acceptable	4.040	.631
Reasonably acceptable	2.294	.722
Barely acceptable	1.078	.518
Neutral	.000	.000
Barely unacceptable	-1.100	.300
Rather unacceptable	-2.020	.836
Substantially unacceptable	-3.235	.899
Highly unacceptable	-4.220	.576
Completely unacceptable	-4.900	.361

From: Matthews, Wright, and Yudowitch (1975). See Section VIII-A 6.

Table VIII-B-4

A second Sample List of Phrases Denoting Degrees of Acceptability

Phrase	Mean	SD
Very, very acceptable	4.157	.825
Largely acceptable	3.137	.991
Mildly acceptable	1.686	.700
Sort of acceptable	.940	.645
Neutral	.000	.000
Barely unacceptable	-1.100	.300
Rather unacceptable	-2.020	.836
Substantially unacceptable	-3.235	.899
Highly unacceptable	-4.294	.535
Completely unacceptable	-4.900	.361

From: Matthews, Wright, and Yudowitch (1975). See Section VIII-A 6.

While the scale values from the studies cited are useful, further refinement is possible. That is, once having selected a candidate scale (set of descriptors) one could then conduct another study to determine if relevant judges would assign scale values indicating equal intervals (among means) for the terms on the candidate scale.

6. Selecting Descriptors for End Points

Once the decision has been made to how extreme the endpoints of a scale should be (see Section VII-E 4), the descriptors should be selected accordingly. If extreme end points are desired, descriptors that have extreme meaning should be selected. One guideline that can be used in selecting these descriptors is to use those that have the highest and lowest scale values. Another guideline is to review the descriptors in terms of their apparent meanings. If less extreme end points are desired, descriptors that do not have extreme scale values and that do not have the apparent extreme meanings should be selected.

7. Selecting Midpoint Responses

In selecting a descriptor for a midpoint response, it is necessary to use a descriptor that is neutral in meaning. Some of the commonly used midpoints do not appear as neutral as might be expected to some respondents.

Table VIII-B-5 lists several neutral terms with their scale values and standard deviations. This list may be helpful in selecting midpoint responses.

Words commonly used for midpoint responses are discussed below:

a. Average.

"Average" should never be used in conjunction with adjectives such as "Excellent," "Good," etc. "Average" has no meaning when used with these words. For example, "Average performance may be superior or it may be completely unsatisfactory. Furthermore, most evaluators do not have the experience or competence to even know what an "average" performance is. Typically, when "Average" is used on a field test evaluation form only 5% or 10% of responders rate the subject as below average and 30% or 40% rate it above average. The data from such a question indicate that the response alternatives are not well formulated. Therefore, as a general rule, it is usually inappropriate to use any term of "Average" in a questionnaire, and it is always inappropriate to use "Average" in conjunction with phrases such as "Excellent," "Good," "Poor," etc.

Table VIII-B-5

Neutral Term Scale Values and Standard Deviations as Determined by Several Different Studies

Term	Mean Scale Value	SD	Theoretical Neutral Scale Value
About average	3.77	.85	3.50
Acceptable	.73	.66	.00
Acceptable	11.12	2.59	10.00
Acceptable	2.39	1.46	.00
All right	10.76	1.42	10.00
Average	3.08	--	3.00
Average	.86	1.08	.00
Average	10.84	1.55	10.00
Borderline	-.02	.32	.00
Borderline	.00	.20	.00
Borderline	-.06	.31	.00
Doesn't make any difference	2.83	3.73[a]	5.00
Don't know	4.82	.82[a]	5.00
Fair	6.50	--	5.50
Fair	.78	.85	.00
Fair	9.52	2.06	10.00
Fair	4.96	.77[a]	5.00
Neutral	.00	.00	.00
Neutral	.02	.18	.00
Neutral	9.80	1.50	10.00
Neutral	10.18	2.01	10.00
Normal	6.70	1.43	6.00
Ordinary	6.50	1.43	6.00
O.K.	.87	1.24	.00
O.K.	10.28	1.67	10.00
So-so	10.08	1.87	10.00
Undecided	4.76	3.73[a]	5.00

[a] Interquartile range shown rather than the standard deviation

If "Average" is used, it should be with extreme care and only when one is interested in comparing performances or items with each other. It should not be used when one desires to find out how "good" or how "bad" an item or performance is. Significantly above average performance may be extremely unsatisfactory.

b. No opinion.

"No opinion" is unacceptable as a neutral term, as it usually denotes that a person has no opinion due to lack of knowledge or due to not having thought about an issue. "No opinion" can be used as a response alternative if it represents a specific type of information that is wanted.

c. Neutral.

"Neutral" is considered as a less desirable term to use than "Borderline." Although every respondent in the study gave the term zero, the meaning on a questionnaire is not clear (see page VIII-B 4). Two out of 52 respondents indicated it was unrateable. In another study "Neutral" had a mean scale value of .02 and a standard deviation of .18. Because of the ambiguity of meaning of "neutral" (e.g., feeling of the respondent versus midpoint alternative) it is not recommended that it be used as mid-point on most questionnaires.

d. Marginal.

"Marginal" is sometimes used as a midpoint response alternative. Interviews with test subjects indicated that the term "Marginal" in most cases had a meaning of above "Borderline" or still satisfactory, but very close to being unsatisfactory. Hence, indications are that there may be more desirable terms to use than "Marginal."

e. Borderline.

"Borderline" is preferred by some experts as a midpoint response. In an administration to Fort Hood soldiers of over 1,500 questionnaires using the term "Borderline" as a midpoint, there was not one instance of reported confusion among those completing the questionnaires. However, there are times when "Borderline" has a larger standard deviation than "Neutral." (Again, "neutral" by definition implies zero to most persons, but it's frame of reference is ambiguous).

f. Uncertain.

"Uncertain" is unacceptable as a neutral term as it implies that with additional knowledge or thought a decision could be made that would fall into one of the other categories.

g. Undecided.

"Undecided" is also unacceptable as a neutral item for the same reasons as "Uncertain."

h. Neither agree nor disagree.

"Neither agree nor disagree" and similar descriptors written in this form may be used as midpoint responses. They have the advantage of paralleling the rest of the descriptors in the set, and they denote a position exactly in the middle of the end points. This term, like "neutral," can also imply uncertainty, indecision or a lack of knowledge rather than a firm knowledge that it represents a mid-point.

i. No effect.

"No effect" may be employed as a netural term when it is used with a set of descriptors to measure the type of effect that an activity will have. For instance, it can be used on a continuum from beneficial to detrimental.

j. Ordinary.

"Ordinary" should not be used as a neutral item. In one study its scale value showed marked skewing at the low extreme, indicative of the common use of "ordinary" to imply inferiority.

k. Fair.

"Fair" should not be used as a neutral item. In one study the median scale value for "fair" was a full point above the neutral point. It appears for some subjects that the meaning of "fair" is distinctly favorable.

l. Acceptable.

"Acceptable" is not a desirable word to use as a neutral item. In one study it exhibited a marked bimodality of response, indicating that subjects disagreed on the degree of favorableness noted by the term. In a recent study "Acceptable" had a large standard deviation of 1.46.

m. Normal.

"Normal" is not a desirable word to use as a neutral item. In one study it exhibited a marked bimodality of response, indicating that the word "normal" has different meanings for different subjects. This term would be classified as a synonym for "average."

n. Medium.

"Medium" may possibly be used as a neutral term. In one study there was a piling up of judgments for "Medium" at the neutral scale position.

o. O.K. or all right.

"O.K." or "All right" has been used sometimes as a midpoint response alternatives. However, they have a tendency to be rated more positively than neutral. They also have larger standard deviations than other terms mentioned, indicating that there is ambiguity in their meaning.

p. So-so.

"So-so" is another term sometimes used as a midpoint response. In one study it had a scale value of 10.08, which was very close to the neutral scale value of 10.00, but it also had a fairly large standard deviation of 1.87. It's use is not recommended.

q. Don't know.

"Don't know" is an unacceptable term to use as a middle point. It usually means to the subject that with additional knowledge or more time to think about the issue, he could choose one of the other alternatives.

r. Doesn't make any difference.

"Doesn't make any difference" should not be used as a midpoint response alternative because it implies a more negative value than a neutral value. In one study it had a scale value of 2.83, where the neutral scale value was 5.00. It also had an interquartile range of 3.13, which means that there was a lot of disagreement among subjects as to its meaning.

8. Selecting Positive and Negative Descriptors

If a balanced scale is desired, it is necessary to select an equal number of positive and negative descriptors. In most cases it is easy to determine if a descriptor is positive or negative by seeing on which side of the neutral point its scale value falls. For example, "Mildly like" has a positive scale value, and "Mildly dislike" has a negative scale value.

9. Selecting Terms Showing Equal Intervals

Some experts argue that, in order to perform analyses on the basis of numerical values or weights, the intervals between rating scale response alternatives should be equal. This would be desirable, but in many cases it is impossible because many words have not been assigned scale values. But when scale values are available, the response alternatives can be selected as equally distant apart as possible when doing so is considered important.

There is a tendency for some questionnaire constructors to select phrases with parallel wording to indicate equal intervals. (They may also do so for other reasons.) However, if equal intervals are considered important, phrases should be selected based upon scale values if available. For example, in Table VIII-E-9 "Highly adequate" has a scale value of 3.843 while the parallel term "Highly inadequate" has a scale value of -4.196. This places "Highly inadequate" further away from the neutral point than "Highly adequate."

10. Use of Unscaled Terms

Some discussion is in order regarding the use of terms ignoring their scale values or to which no scale values have been assigned. An illustration of the first of these practices is from a study in which ARI had 21 Army officers involved in operational field testing rank-order 16 terms that included "Outstanding," "Superior," "Excellent" and "Very Good." "Excellent" was ranked as less positive than "Outstanding" by 14 of the officers, while it was ranked as less positive than "Superior" by 17 of the officers. However, there was maximum disagreement as to whether "Outstanding" or "Superior" was first or second on the scale. That is, 12 rated "Superior" first and "Outstanding" second, while nine of the officers assigned the reverse ordering to these two words. All officers ranked "Outstanding," "Superior," and "Excellent" as more positive than "Very Good." "Outstanding" is sometimes interpreted to denote only that the performance is among the best of a group - without any implication as to quality, e.g., although a student's grade of 65 out of 100 points was failing, his performance may have been "Outstanding" since no other student in the class scored above 60!

What are the consequences to the developer of rating scale items of discovering a mean 50%-50% split as in the ordering of "Outstanding" and "Superior"? Does it mean they *cannot* be used together as part of the descriptors of a rating scale item? The answer is, "Normally yes." In Figure VIII-B-1, we would have better discrimination if "Outstanding" were replaced by "Excellent," with the position formerly occupied by "Excellent" being filled by "Very good." "Superior" and "Outstanding" or similarly overlapping terms should normally not be used on the same scale.

Figure VIII-B-1

Two Formats Using "Outstanding" and "Superior"

1. _____ 1. Superior
_____ 2. Outstanding
_____ 3. Excellent
_____ 4. Good
_____ 5. Fair
_____ 6. Poor

2. Superior Outstanding Excellent Good Fair Poor

____
(Circle one Word)

When functioning as questionnaire consultants or developers in field test situations where respondents are enlisted personnel ARI has recommended and used very little variety in its rating scale items. Arrays such as those shown in Figure VIII-B-2 are almost always proposed and used. Sometimes the middle term is deleted. Several reasons for the lack of variety are that a standard simple format 1) facilitates comparability of rating distributions with previous tests, and 2) facilitates understanding by soldier respondents, who are often not high school graduates.

Figure VIII-B-2

Response Alternatives
Frequently Recommended by ARI

( ) Very satisfactory

( ) Satisfactory

( ) Borderline

( ) Unsatisfactory

( ) Very unsatisfactory

( ) Very effective

( ) Effective

( ) Borderline

( ) Ineffective

( ) Very ineffective

( ) Very acceptable

( ) Acceptable

( ) Borderline

( ) Unacceptable

( ) Very unacceptable

C. Selection of Response Alternatives Denoting Degrees of Frequency

Some questionnaire designers use verbal descriptors to denote degrees of frequency. Table VIII-C-1 shows such a list of verbal descriptors. A study showed that there was a great deal of variability in meaning for frequency phrases. Questionnaires should, whenever possible, use response alternatives that include a number designation or percentage of time meant by each word used as a response alternative.

Table VIII-C-1

Degrees of Frequency

Phrase	Scale Value	Inter-Quartile Range
Always	8.99	.52
Without fail	8.89	.61
Often	7.23	1.02
Usually	7.17	1.36
Frequently	6.92	.77
Now and then	4.79	1.40
Sometimes	4.78	1.83
Occasionally	4.13	2.06
Seldom	2.45	1.05
Rarely	2.08	.61
Never	1.00	.50

From: Dodd and Gerberick (1960). See Section VIII-A 3.

D. Selection of Response Alternatives Using Order of Merit Lists of Descriptor Terms

An order of merit list of descriptors does not provide scale values nor show the variance of each phrase of some continuum. In addition, the list does not represent an equal interval scale. However, such lists are still useful for selecting response alternatives, if the main concern is to select response categories so that each respondent will agree on the relative degree of "goodness" of the terms. Tables VIII-D-1 and VIII-D-2 give examples of order of merit lists of descriptor terms.

Table VIII-D-1

Order of Merit of Selected Descriptive Terms

Order of merit	Descriptive Term
1	Very superior
2	Very outstanding
3	Superior
4	Outstanding
5	Excellent
6	Very good
7	Good
8	Very satisfactory
9	Satisfactory
10	Marginal
11	Borderline
12	Poor
13	Unsatisfactory
14	Bad
15	Very poor
16	Very unsatisfactory
17	Very bad
18	Extremely poor
19	Extremely unsatisfactory
20	Extremely bad

From: Gividen (1973). Section VIII-A 4.

Table VIII-D-2

Order of Merit of Descriptive Terms
Using "Use" as a Descriptor

Order of merit	Descriptive term
1	Extremely useful
2	Very useful
3	Of significant use
4	Of considerable use
5	Of much use
6	Of moderate use
7	Of use
8	Of some use
9	Of little use
10	Not very useful
11	Of slight use
12	Of very little use
13	Of no use

From: Gividen (1973). See Section VIII-A 4.

E. Selection of Response Alternatives Using Scale Values and Standard Deviations

Using scale values and standard deviations to select response alternatives will give a more refined set of phrases than using an order of merit list. Other sections above have discussed specific considerations in selecting descriptors. In general, response alternatives selected from lists of phrases with scale values should usually have the following characteristics:

1. The scale vaules of the terms should be as far apart as possible.
2. The scale values of the terms should be as equally distant as possible.
3. The terms should have small variability (small standard deviations or interquartile ranges).
4. Other things being equal, the terms should have parallel wording.

Tables VIII-E-1 through VIII-E-24 give lists of phrases which have scale values and, when possible, standard deviations or interquartile range. They are based on empirical evidence, and may be used to select response alternatives.

Table VIII-E-1

Acceptability Phrases

Phrase	Average	SD
Excellent	6.27	.54
Perfect in every respect	6.22	.86
Extremely good	5.74	.81
Very good	5.19	.75
Unusually good	5.03	.98
Very good in most respects	4.62	.72
Good	4.25	.90
Moderately good	3.58	.77
Could use some minor changes	3.28	1.09
Not good enough for extreme conditions	3.10	1.30
Not good for rough use	2.72	1.15
Not very good	2.10	.85
Needs major changes	1.97	1.12
Barely acceptable	1.79	.90
Not good enough for general use	1.76	1.21
Better than nothing	1.22	1.08
Poor	1.06	1.11
Very poor	.76	.95
Extremely poor	.36	.76

From: U.S. Army (1973). See Section VIII-A 9.

Table VIII-E-2

Degrees of Excellence: First Set

Phrase	Scale Value	SD
Superior	20.12	1.17
Fantastic	20.12	0.83
Tremendous	19.84	1.31
Superb	19.80	1.19
Excellent	19.40	1.73
Terrific	19.00	2.45
Outstanding	18.96	1.99
Wonderful	17.32	2.30
Delightful	16.92	1.85
Fine	14.80	2.12
Good	14.32	2.08
Pleasant	13.44	2.06
Nice	12.56	2.14
Acceptable	11.12	2.59
Average	10.84	1.55
All right	10.76	1.42
O.K.	10.28	1.67
Neutral	9.80	1.50
Fair	9.52	2.06
Mediocre	9.44	1.80
Unpleasant	5.04	2.82
Bad	3.88	2.19
Very bad	3.20	2.10
Unacceptable	2.64	2.04
Awful	1.92	1.50
Terrible	1.76	.77
Horrible	1.48	.87

From: Myers and Warner (1968). See Section VIII-A 8.

Table VIII-E-3

Degrees of Excellence: Second Set

Phrase	Scale Value	SD
Best of all	6.15	2.48
Excellent	3.71	1.01
Wonderful	3.51	.97
Mighty fine	2.88	.67
Especially good	2.86	.82
Very good	2.56	.87
Good	1.91	.76
Pleasing	1.58	.65
O.K.	.87	1.24
Fair	.78	.85
Only fair	.71	.64
Not pleasing	-.83	.67
Poor	-1.55	.87
Bad	-2.02	.80
Very bad	-2.53	.64
Terrible	-3.09	.98

From: Jones and Thurstone (1955).
See Section VIII-A 5.

Table VIII-E-4

Degrees of Like and Dislike

Phrase	Scale Value	SD
Like extremely	4.16	1.62
Like intensely	4.05	1.59
Strongly like	2.96	.69
Like very much	2.91	.60
Like very well	2.60	.78
Like quite a bit	2.32	.52
Like fairly well	1.51	.59
Like	1.35	.77
Like moderately	1.12	.61
Mildly like	.85	.47
Like slightly	.69	.32
Neutral	.02	.18
Like not so well	-.30	1.07
Like not so much	-.41	.94
Dislike slightly	-.59	.27
Mildly dislike	-.74	.35
Dislike moderately	-1.20	.41
Dislike	-1.58	.94
Don't like	-1.81	.97
Strongly dislike	-2.37	.53
Dislike very much	-2.49	.64
Dislike intensely	-3.33	1.39
Dislike extremely	-4.32	1.86

From: Jones and Thurstone (1955).
See Section VIII-A 5.

Table VIII-E-5

Degrees of Good and Poor

Phrase	Scale Value	SD
Exceptionally good	18.56	2.36
Extremely good	18.44	1.61
Unusually good	17.08	2.43
Remarkably good	16.68	2.19
Very good	15.44	2.77
Quite good	14.44	2.76
Good	14.32	2.08
Moderately good	13.44	2.23
Reasonably good	12.92	2.93
Fairly good	11.96	2.42
Slightly good	11.84	2.19
So-so	10.08	1.87
Not very good	6.72	2.82
Moderately poor	6.44	1.64
Reasonably poor	6.32	2.46
Slightly poor	5.92	1.96
Poor	5.72	2.09
Fairly poor	5.64	1.68
Quite poor	4.80	1.44
Unusually poor	3.20	1.44
Very poor	3.12	1.17
Remarkably poor	2.88	1.74
Exceptionally poor	2.52	1.19
Extremely poor	2.08	1.19

From: Myers and Warner (1968).
See Section VIII-A 8.

Table VIII-E-6

Degrees of Good and Bad

Phrase	Scale Value
Extremely good	3.449
Very good	3.250
Unusually good	3.243
Decidedly good	3.024
Quite good	2.880
Rather good	2.755
Good	2.712
Pretty good	2.622
Somewhat good	2.462
Slightly good	2.417
Slightly bad	1.497
Somewhat bad	1.323
Rather bad	1.232
Bad	1.024
Pretty bad	1.018
Quite bad	.924
Decidedly bad	.797
Unusually bad	.662
Very bad	.639
Extremely bad	.470

From: Cliff (1959). See Section VIII-A 2.

Table VIII-E-7

Degrees of Agree and Disagree

Phrase	Mean	SD
Decidedly agree	2.77	.41
Quite agree	2.37	.49
Considerably agree	2.21	.42
Substantially agree	2.10	.50
Moderately agree	1.47	.41
Somewhat agree	.94	.41
Slightly agree	.67	.36
Perhaps agree	.52	.46
Perhaps disagree	-.43	.46
Slightly disagree	-.64	.38
Somewhat disagree	-.93	.47
Moderately disagree	-1.35	.42
Quite disagree	-2.16	.57
Substantially disagree	-2.17	.51
Considerably disagree	-2.17	.45
Decidedly disagree	-2.76	.43

From: Altemeyer (1970). See Section VIII-A 1.

Table VIII-E-8

Degrees of More and Less

Phrase	Scale Value	Inter-quartile Range[a]
Very much more	8.02	.61
Much more	7.67	1.04
A lot more	7.50	1.06
A good deal more	7.29	.98
More	6.33	1.01
Somewhat more	6.25	.98
A little more	6.00	.58
Slightly more	5.99	.57
Slightly less	3.97	.56
A little less	3.96	.54
Less	3.64	1.04
Much less	2.55	1.06
A good deal less	2.44	1.11
A lot less	2.36	1.03
Very much less	1.96	.52

From: Dodd and Gerberick (1960).
See Section VIII-A 3.

[a] Minimum = 0.5.

Table VIII-E-9

Degrees of Adequate and Inadequate

Phrase	Mean	SD
Totally adequate	4.620	.846
Absolutely adequate	4.540	.921
Completely adequate	4.490	.825
Extremely adequate	4.412	.719
Exceptionally adequate	4.380	.869
Entirely adequate	4.340	.863
Wholly adequate	4.314	1.038
Fully adequate	4.294	.914
Very very adequate	4.063	.876
Perfectly adequate	3.922	1.026
Highly adequate	3.843	.606
Most adequate	3.843	.978
Very adequate	3.420	.851
Decidedly adequate	3.140	1.536
Considerably adequate	3.020	.874
Quite adequate	2.980	.979
Largely adequate	2.863	.991
Substantially adequate	2.608	1.030
Reasonably adequate	2.412	.771
Pretty adequate	2.306	.862
Rather adequate	1.755	.893
Mildly adequate	1.571	.670
Somewhat adequate	1.327	.793
Slightly adequate	1.200	.566
Barely adequate	.627	.928
Neutral	.000	.000
Borderline	-.020	.316
Barely inadequate	-1.157	.638
Mildly inadequate	-1.353	.621
Slightly inadequate	-1.380	.772
Somewhat inadequate	-1.882	.732
Rather inadequate	-2.102	.974
Moderately inadequate	-2.157	1.017
Fairly inadequate	-2.216	.800
Pretty inadequate	-2.347	.959
Considerably inadequate	-3.600	.680
Very inadequate	-3.735	.777
Decidedly inadequate	-3.780	.944
Most inadequate	-3.980	1.545
Highly inadequate	-4.196	.741

(Table continued on next page)

Table VIII-E-9 (Cont.)

Degrees of Adequate and Inadequate

Phrase	Mean	SD
Very very inadequate	-4.460	.537
Extremely inadequate	-4.608	.527
Fully inadequate	-4.667	.676
Exceptionally inadequate	-4.680	.508
Wholly inadequate	-4.784	.498
Entirely inadequate	-4.792	.644
Completely inadequate	-4.800	.529
Absolutely inadequate	-4.880	.431
Totally inadequate	-4.900	.412

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 6.

Table VIII-E-10

Degrees of Acceptable and Unacceptable

Phrase	Mean	SD
Wholly acceptable	4.725	.563
Completely acceptable	4.686	.610
Fully acceptable	4.412	..867
Extremely acceptable	4.392	.716
Most acceptable	4.157	.915
Very very acceptable	4.157	.825
Highly acceptable	4.040	.631
Quite acceptable	3.216	.956
Largely acceptable	3.137	.991
Acceptable	2.392	1.456
Reasonably acceptable	2.294	.722
Moderately acceptable	2.280	.722
Pretty acceptable	2.000	1.125

(Table continued on next page)

Table VIII-E-10 (Cont.)

Degrees of Acceptable and Unacceptable

Phrase	Mean	SD
Rather acceptable	1.939	.818
Fairly acceptable	1.840	.924
Mildly acceptable	1.686	.700
Somewhat acceptable	1.458	1.241
Barely acceptable	1.078	.518
Slightly acceptable	1.039	.522
Sort of acceptable	.940	.645
Borderline	.000	.200
Neutral	.000	.000
Marginal	-.120	.515
Barely unacceptable	-1.100	.300
Slightly unacceptable	-1.255	.589
Somewhat unacceptable	-1.765	.674
Rather unacceptable	-2.020	.836
Fairly unacceptable	-2.160	.880
Moderately unacceptable	-2.340	.681
Pretty unacceptable	-2.412	.662
Reasonably unacceptable	-2.440	.753
Unacceptable	-2.667	1.381
Substantially unacceptable	-3.235	.899
Quite unacceptable	-3.388	1.066
Largely unacceptable	-3.392	.818
Considerably unacceptable	-3.440	.779
Notably unacceptable	-3.500	1.044
Decidedly unacceptable	-3.837	1.017
Highly unacceptable	-4.294	.535
Most unacceptable	-4.420	.724
Very very unacceptable	-4.490	.500
Exceptionally unacceptable	-4.540	.607
Extremely unacceptable	-4.686	.464
Completely unacceptable	-4.900	.361
Entirely unacceptable	-4.900	.361
Wholly unacceptable	-4.922	.269
Absolutely unacceptable	-4.922	.334
Totally unacceptable	-4.941	.235

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 6.

Table VIII-E-11

Comparison Phrases

Phrase	Mean	SD
Best of all	4.896	.510
Absolutely best	4.843	.459
Truly best	4.600	.721
Undoubtedly best	4.569	.823
Decidedly best	4.373	.839
Best	4.216	1.459
Absolutely better	4.060	.988
Extremely better	3.922	.882
Substantially best	3.700	.922
Decidedly better	3.412	.933
Conspicuously better	3.059	.802
Moderately better	2.255	.737
Somewhat better	1.843	.801
Rather better	1.816	.719
Slightly better	1.157	.776
Barely better	.961	.656
Absolutely alike	.588	1.623
Alike	.216	.847
The same	.157	.801
Neutral	.000	.000
Borderline	-.061	.314
Marginal	-.184	.919
Barely worse	-1.039	.816
Slightly worse	-1.216	.498
Somewhat worse	-2.078	.860
Moderately worse	-2.220	.944
Noticeably worse	-2.529	1.036
Worse	-2.667	1.423
Notably worse	-3.020	1.038
Largely worse	-3.216	1.108
Considerably worse	-3.275	1.206
Conspicuously worse	-3.275	.887
Much worse	-3.286	.808
Substantially worse	-3.460	.899
Decidedly worse	-3.760	.907
Very much worse	-3.941	.752
Absolutely worse	-4.431	.823
Decidedly worst	-4.431	.748
Undoubtedly worst	-4.510	.872
Absolutely worst	-4.686	1.291
Worst of all	-4.776	1.298

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 6.

Table VIII-E-12

Degrees of Satisfactory and Unsatisfactory

Phrase	Scale Value	SD
Quite satisfactory	4.35	.95
Satisfactory	3.69	.87
Not very satisfactory	2.11	.76
Unsatisfactory but usable	2.00	.87
Very unsatisfactory	.69	1.32

From: U.S. Army (1973). See Section VIII-A 9.

Table VIII-E-13

Degrees of Unsatisfactory

Phrase	Scale Value
Unsatisfactory	1.47
Quite unsatisfactory	1.00
Very unsatisfactory	.75
Unusually unsatisfactory	.75
Highly unsatisfactory	.71
Very, very unsatisfactory	.25
Extremely unsatisfactory	.10
Completely unsatisfactory	.00

From: Mosier (1941). See Section VIII-A 7.

Table VIII-E-14

Degrees of Pleasant

Phrase	Scale Value
Extremely pleasant	3.490
Very pleasant	3.174
Unusually pleasant	3.107
Decidedly pleasant	3.028
Quite pleasant	2.849
Pleasant	2.770
Rather pleasant	2.743
Pretty pleasant	2.738
Somewhat pleasant	2.505
Slightly pleasant	2.440

From: Cliff (1959). See Section VIII-A 2.

Table VIII-E-15

Degrees of Agreeable

Phrase	Scale Value
Very, very agreeable	5.34
Extremely agreeable	5.10
Highly agreeable	5.02
Completely agreeable	4.96
Unusually agreeable	4.86
Very agreeable	4.82
Quite agreeable	4.45
Agreeable	4.19

From: Mosier (1941). See Section VIII-A 7.

Table VIII-E-16

Degrees of Desirable

Phrase	Scale Value
Very, very desirable	5.66
Extremely desirable	5.42
Completely desirable	5.38
Unusually desirable	5.23
Highly desirable	5.15
Very desirable	4.96
Quite desirable	4.76
Desirable	4.50

From: Mosier (1941). See Section VIII-A 7.

Table VIII-E-17

Degrees of Nice

Phrase	Scale Value
Extremely nice	3.351
Unusually nice	3.155
Very nice	3.016
Decidedly nice	2.969
Pretty nice	2.767
Quite nice	2.738
Nice	2.636
Rather nice	2.568
Somewhat nice	2.488
Slightly nice	2.286

From: Cliff (1959). See Section VIII-A 2.

Table VIII-E 18

Degrees of Adequate

Phrase	Scale Value	SD
More than adequate	4.13	1.11
Adequate	3.39	.87
Not quite adequate	2.40	.85
Barely adequate	2.10	.84
Not adequate	1.83	.98

From: U.S. Army (1973). See Section VIII-A 9.

Table VIII-E-19

Degrees of Ordinary

Phrase	Scale Value
Ordinary	2.074
Very ordinary	2.073
Somewhat ordinary	2.038
Rather ordinary	2.034
Pretty ordinary	2.026
Slightly ordinary	1.980
Decidedly ordinary	1.949
Extremely ordinary	1.936
Unusually ordinary	1.875

From: Cliff (1959). See Section VIII-A 2.

Table VIII-E-20

Degrees of Average

Phrase	Scale Value
Rather average	2.172
Average	2.145
Quite average	2.101
Pretty average	2.094
Somewhat average	2.080
Unusually average	2.062
Extremely average	2.052
Very average	2.039
Slightly average	2.023
Decidedly average	2.020

From: Cliff (1959). See Section VIII-A 2.

Table VIII-E-21

Degrees of Hesitation

Phrase	Scale Value	Inter-quartile Range[a]
Without hesitation	7.50	6.54
With little hesitation	5.83	3.40
Hesitant	4.77	1.06
With some hesitation	4.38	1.60
With considerable hesitation	3.29	3.39
With much hesitation	3.20	5.25
With great hesitation	2.41	6.00

From: Dodd and Gerberick (1960). See Section VIII-A 3.

[a] Minimum = 0.5.

Table VIII-E-22

Degrees of Inferior

Phrase	Scale Value
Slightly inferior	1.520
Somewhat inferior	1.516
Inferior	1.323
Rather inferior	1.295
Pretty inferior	1.180
Quite inferior	1.127
Decidedly inferior	1.013
Unusually inferior	.963
Very inferior	.927
Extremely inferior	.705

From: Cliff (1959). See Section VIII-A 2.

Table VIII-E-23

Degrees of Poor

Phrase	Scale Value
Poor	1.60
Quite poor	1.30
Very poor	1.18
Unusually poor	.95
Extremely poor	.95
Completely poor	.92
Very, very poor	.55

From: Mosier (1941). See Section VIII-A 7.

Table VIII-E-24

Descriptive Phrases

Phrase	Scale Value	Inter-quartile Range[a]
Complete	8.85	.65
Extremely vital	8.79	.84
Very certain	8.55	1.05
Very strongly	8.40	1.04
Very crucial	8.29	1.12
Very important	8.22	1.16
Very sure	8.15	.95
Almost complete	8.06	.58
Of great importance	8.05	.91
Very urgent	8.00	.90
Feel strongly toward	7.80	1.60
Essential	7.58	1.85
Very vital	7.55	1.05
Certain	7.13	1.44
Strongly	7.07	.67
Important	6.83	1.14
Good	6.72	1.20
Urgent	6.41	1.53
Crucial	6.39	1.73
Sure	5.93	1.87
Vital	5.92	1.63
Moderately	5.24	.99
Now	5.03	.53
As at present	5.00	.50
Fair	4.96	.77
Don't know	4.82	.82
Undecided	4.76	1.06
Don't care	4.63	2.00
Somewhat	3.79	.94
Indifferent	3.70	2.20
Object strongly to	3.50	6.07
Not important	3.09	1.33
Unimportant	1.94	1.42
Bad	2.83	.93
Uncertain	2.83	2.50
Doesn't make any difference	2.83	3.13
Not sure	2.82	1.24
Not certain	2.64	2.62

(Table continued on next page)

Table VIII-E-24 (Cont.)

Descriptive Phrases

Phrase	Scale Value	Inter-quartile Range[a]
Non-essential	2.58	1.67
Doesn't mean anything	2.50	2.71
Insignificant	2.12	1.14
Very little	2.08	.64
Almost none	2.04	.57
Very unimportant	1.75	1.25
Only as a last resort	1.70	7.30
Very bad	1.50	1.13
None	1.11	.59

From; Dodd and Gerberick (1960). See Section VIII-A 3.

[a] Minimum = 0.5.

F. Sample Sets of Response Alternatives

It is sometimes valuable and is a time saver to have lists of response alternatives available to use. The tables in this section give some examples of response alternatives that have been selected on different bases. These sets do not exhaust all possibilities.

The sets of response alternatives that appear in Table VIII-F-1 were selected so that the phrases in each set would have means at least one standard deviation away from each other and have parallel wording. Some of the sets of response alternatives have extreme end points, some do not. The sets of response alternatives shown in Table VIII-F-2 were selected so that the phrases in each set would be as nearly equally distant from each other as possible without regard to parallel wording. Table VIII-F-3 contains sets of response alternatives selected from lists of descriptors with only scale values given. The phrases were selected on the bases of equal appearing intervals. Table VIII-F-4 has sets of response alternatives selected from order of merit lists of descriptors.

Table VIII-F-1

Sets of Response Alternatives Selected so Phrases Are at Least One Standard Deviation Apart and Have Parallel Wording

Set No.	Response Alternatives
1.	Completely acceptable Reasonably acceptable Barely acceptable Borderline Barely unacceptable Reasonably unacceptable Completely unacceptable
2.	Wholly acceptable Largely acceptable Borderline Largely unacceptable Wholly unacceptable
3.	Largely acceptable Barely acceptable Borderline Barely unacceptable Largely unacceptable
4.	Reasonably acceptable Slightly acceptable Borderline Slightly unacceptable Reasonably unacceptable
5.	Totally adequate Very adequate Barely adequate Borderline Barely inadequate Very inadequate Totally inadequate
6.	Completely adequate Considerably adequate Borderline Considerably inadequate Completely inadequate
7.	Very adequate Slightly adequate Borderline Slightly inadequate Very inadequate
8.	Highly adequate Mildly adequate Borderline Mildly inadequate Highly inadequate
9.	Decidedly agree Substantially agree Slightly agree Slightly disagree Substantially disagree Decidedly disagree
10.	Moderately agree Perhaps agree Neutral Perhaps disagree Moderately disagree
11.	Undoubtedly best Conspicuously better Moderately better Alike Moderately worse Conspicuously worse Undoubtedly worst
12.	Moderately better Barely better The same Barely worse Moderately worse

(Table continued on next page)

Table VIII-F-1 (Cont.)

Sets of Response Alternatives Selected so Phrases Are at Least One Standard Deviation Apart and Have Parallel Wording

Set No.	Response Alternatives	Set No.	Response Alternatives
13.	Extremely good	16.	Like extremely
	Remarkably good		Like moderately
	Good		Neutral
	So-so		Dislike moderately
	Poor		Dislike extremely
	Remarkably poor		
	Extremely poor	17.	Strongly like
			Like
14.	Exceptionally good		Neutral
	Reasonably good		Don't like
	So-so		Strongly dislike
	Reasonably poor		
	Exceptionally poor	18.	Very much more
			A good deal more
15.	Very important		A little more
	Important		A little less
	Not important		A good deal less
	Very unimportant		Very much less

Table VIII-F-2

Sets of Response Alternatives Selected so That Intervals Between Phrases Are as Nearly Equal as Possible

Set No.	Response Alternatives
1.	Completely acceptable
	Reasonably acceptable
	Borderline
	Moderately unacceptable
	Extremely unacceptable
2.	Totally adequate
	Pretty adequate
	Borderline
	Pretty inadequate
	Extremely inadequate
3.	Highly adequate
	Rather adequate
	Borderline
	Somewhat inadequate
	Decidedly inadequate
4.	Quite agree
	Moderately agree
	Perhaps agree
	Perhaps disagree
	Moderately disagree
	Substantially disagree
5.	Undoubtedly best
	Moderately better
	Borderline
	Noticeably worse
	Undoubtedly worst
6.	Fantastic
	Delightful
	Nice
	Mediocre
	Unpleasant
	Horrible
7.	Perfect in every respect
	Very good
	Good
	Could use some minor changes
	Not very good
	Better than nothing
	Extremely poor
8.	Excellent
	Good
	Only fair
	Poor
	Terrible
9.	Extremely good
	Quite good
	So-so
	Slightly poor
	Extremely poor
10.	Remarkably good
	Moderately good
	So-so
	Not very good
	Unusually poor
11.	Without hesitation
	With little hesitation
	With some hesitation
	With great hesitation
12.	Strongly like
	Like quite a bit
	Like
	Neutral
	Mildly dislike
	Dislike very much
	Dislike extremely

(Table continued on next page)

Table VIII-F-2 (Cont.)

Sets of Response Alternatives Selected so That Intervals Between Phrases Are as Nearly Equal as Possible

Set No.	Response Alternatives	Set No.	Response Alternatives
13.	Like quite a bit Like Like slightly Borderline Dislike slightly Dislike moderately Don't like	15.	Very much more A little more Slightly less Very much less
14.	Like quite a bit Like fairly well Borderline Dislike moderately Dislike very much		

Table VIII-F-3

Sets of Response Alternatives Selected from Lists Giving Scale Values Only

Set No.	Response Alternatives	Set No.	Response Alternatives
1.	Very, very agreeable Usually agreeable Quite agreeable Agreeable	6.	Extremely nice Decidedly nice Nice Slightly nice
2.	Rather average Quite average Unusually average Decidedly average	7.	Ordinary Slightly ordinary Unusually ordinary
3.	Very, very desirable Completely desirable Very desirable Desirable	8.	Extremely pleasant Decidedly pleasant Somewhat pleasant
4.	Extremely good Somewhat good Slightly bad Extremely bad	9.	Poor Very poor Very, very poor
5.	Slightly inferior Rather inferior Unusually inferior Extremely inferior	10.	Very, very agreeable Extremely agreeable Very agreeable Quite agreeable Agreeable

Note. Selected so that intervals between phrases are as equal as possible.

Table VIII-F-4

Sets of Response Alternatives Selected Using Order of Merit Lists of Descriptor Terms

Set No.	Response Alternatives
1.	Very good Good Borderline Poor Very poor
2.	Very satisfactory Satisfactory Borderline Unsatisfactory Very unsatisfactory
3.	Very superior Superior Borderline Poor Very poor
4.	Extremely useful Of considerable use Of use Not very useful Of no use

Chapter IX: Physical Characteristics of Questionnaires

A. Overview

This chapter considers four topics related to the physical characteristics of questionnaires: the location of response alternatives relative to the stem (Section IX-B); questionnaire length (Section IX-C); questionnaire format considerations (Section IX-D); and the use of answer sheets (Section IX-E).

B. Location of Response Alternatives Relative to the Stem

Research to determine what effect the location of response alternatives relative to the question stem has on subjects' responses is practically nonexistent. There is some evidence, however, that untrained raters can make relatively error-free graphic ratings regardless of whether the "good" end of the scale is at the left, right, top, or bottom.

In designing a specific questionnaire, the following points should be considered regarding the location of response alternatives relative to the stem:

1. With multiple choice items, the response alternatives are usually arranged vertically under the stem as shown in Section IV-C 2. With a large number of response alternatives, two or more columns of vertically arranged alternatives might be used. Sometimes, if there are only two or three alternatives (such as "Yes" and "No"), they are placed horizontally rather than vertically.

2. Graphic rating scales are usually placed horizontally on a page. However, the descriptive words, phrases, or sentences on a scale should be concentrated as much as possible at specific points on the scale. This is usually easier if the scales are placed vertically on the page, but it can be done either way. Descriptors need not be equally spaced along graphic scales, and should not be if there is reason to believe the psychological distances between them are not equal.

3. With nongraphic (or "numerical") rating scale items and with ranking and forced choice items, the response alternatives are usually placed vertically under the question stem. See examples in Chapter IV. Sometimes rating scale items are placed horizontally under the stem as shown in Section VII-B. If a number of rating scale items all use the same response alternatives, the question stems can be presented in a column with the response alternatives to the right as shown in Figure IX-B-1.

   In Figure IX-B-1 the response alternatives have been rotated 90 degrees to save space. An effort should be made to place the response alternative horizontal with the bottom of the page so that the respondent does not need to turn the page sideways to read them.

4. The response alternatives for semantic differential items are usually placed horizontally on the page. For an example, see Section IV-H.

Figure IX-B-1

Arrangement of Items With Same Rating Scale Response Alternatives

1. How satisfied or dissatisfied are you with each of the following factors or things?

	Very Satisfied	Satisfied	Borderline	Dissatisfied	Very Dissatisfied
a. Type of furniture in barracks.	____	____	____	____	____
b. Medical service to soldiers.	____	____	____	____	____
c. Quality of mess hall food.	____	____	____	____	____
d. Leadership of generals.	____	____	____	____	____
e. Opportunity for promotion.	____	____	____	____	____
f. Army pay.	____	____	____	____	____
g Civilian opinion of Army.	____	____	____	____	____

C. Questionnaire Length

1. General

The length of questionnaires used in field tests has ranged from one page to as many as 30 pages, perhaps more. How long can one expect a respondent to work effectively at the questionnaire-answering task? At what point does attention and motivation start to degrade, thereby producing poorly considered responses or the omission of responses? Research information on this point is not available to provide a basis for a firm recommendation. There is even disagreement on the effect of questionnaire length on the response rate to mailed questionnaires. However, questionnaires which require longer than one hour to complete will, in most situations, cause bordom and indifference. Even 10 or 15 minutes may be too long, if the questionnaire is perceived by the respondent as redundant or asking unnecessary questions. If one is concerned over the effects of a long questionnaire, alternate forms should be used, wherein the order of items is reversed (or approximately so). For example, the items answered last on 50% of the forms would be answered first on the other 50% of the forms. One could also split the respondent group in half and give half of the questions to each group--provided that the two groups were fairly equivalent in relevant characteristics. It is assumed that everything else would already have been done to reduce the number of items before one of these approaches is used.

2. Results of a Recent Study

In a 1976 study, ARI assisted TCATA in obtaining and analyzing questionnaire responses from a group of trainees whose duration and location of basic and advanced individual training was handled differently from the usual. The number of trainees answering items 1-7 and 48-54 of a 54 item questionnaire is shown below. Note that there is very little drop in the number of men in either group as we skip from items 1-7 to items 48-52. This suggests that a 50 item questionnaire, administered as this was, was not so long that persons stopped responding after answering successively more questions.

Now note the sharp drop-about 15% and 9% for the two groups-in responses to items 53 and 54. A more gradual decrease in number of people responding is more what one would expect if they are being "worn down" or fatigued by excessive length.

This result was puzzling, but then it was noted that items 53 and 54 are alone together on the tenth and final page of the questionnaire. It is speculated that many/most of those not answering items 53 and 54 turned page 10 over along with page 9 and thought they had answered all that was required of them. No one checked their questionnaires when they were handed in to see if they had left any items blank. The reductions in respondents appears more of a "last page phenomona" than a consequence of an excessively long questionnaire.

Item #	Experimental Group	Control Group
1	716	512
2	716	513
3	717	511
4	714	513
5	716	514
6	713	510
7	716	511
:	:	:
48	707	509
49	707	508
50	707	508
51	707	510
52	698	505
53	593	462
54	604	461

D. Questionnaire Format Considerations

This section addresses the format of questionnaire items, title and other identification marks, printed introductions, planning to facilitate processing, and other questionnaire format considerations.

1. Format of Questionnaire Items and Format Bias

Item format biases occur when responses to items (questions) are influenced by the question stem or response alternatives. The following guidance is provided:

a. The format of all questionnaire items on a questionnaire should be consistent whenever possible. Mixing multiple choice questions, open ended questions, scales, etc., is normally not desirable.

b. Punctuation and question structure should be consistent and in accordance with proper sentence structure principles. Where incomplete sentences (e.g., "The training that I have received at Fort Hood has been" with five response alternatives of "very challenging" through "very unchallenging") are used as stems no extraneous punctuation, such as a colon, need be put at the end of the stem. The first word of the response alternatives should not be capitalized unless they would be if the statement were written as a continuous sentence. Terminal punctuation at the end of the response alternatives should follow the same general rule of consistency with normal sentence structure. Hence, a period would ordinarily be placed after each response alternative.

When an item consists of a complete question (e.g., "How satisfied or dissatisfied are you with the furniture in the barracks?") the first word of the response alternatives should be capitalized since they do not continue a sentence. If the response alternatives constitute complete sentences, then they should have periods at the end, or whatever other terminal punctuation is appropriate. Sometimes periods are placed at the end of extremely long response alternatives even if they are not sentences. Ordinarily, then, with this form of items, periods would not be placed after the response alternatives.

Exceptions to the above suggestions should be made whenever the exception would improve clarity. An example might be when periods would be confused with decimal points.

c. When items are ambiguous, a recognizable pattern of responses is often produced.

d. Item format bias may be a function of how items are sequenced and grouped.

e. Some authors conclude that a bias can be expected from all closed-ended questions where answers must be selected from two or more fixed choices.

f. The paired comparison format may be useful for those respondents who tend to check many items from a list, and for those who check only a few.

g. Card sorting may show the least item format bias.

h. With two-way choices, some respondents have a tendency to select the first alternative. Others have a tendency to select the second. With other multiple choice items, some respondents have a tendency to select certain categories.

i. There is a little evidence that the first alternative for an item is chosen somewhat more frequently than the others.

2. Title and Other Identification Marks

Each questionnaire should carry a descriptive title centered at the top of the first page of questions and on the instructional and/or introductory cover page if such is used. Each questionnaire form should also be designated by form number to distinguish it from other forms. This number usually goes in the upper left hand corner of each page.

3. Printed Introductions

Introductions are sometimes printed at the start of a questionnaire to tell respondents the purpose and importance of the questionnaire, and the importance of their cooperation in answering all questions carefully. Methodological research is needed to determine the effectiveness of such introductions, but if they are too lengthy there is always the possibility that they might be counterproductive. Regardless, if the introduction is going to run more than a quarter of a page, it might better be placed on a cover sheet.

See Section X-B about questionnaire instructions.

4. Planning to Facilitate Processing

Where possible, questionnaires should be planned to facilitate data collection, processing, and analyses. This frequently involves formulating the questionnaire for machine processing. For small samples, however, manual processing should normally be employed since the effort needed to plan for machine processing is not justified by anticipated data reduction time savings. How to format a questionnaire for machine processing is outside the current scope of this manual. See Section IX-E regarding the use of answer sheets.

5. Other Questionnaire Format Considerations

a. If the respondent's name, rank, etc., is really needed, ask for it on the front page. (See also Section X-C.) Sometimes other information is needed about a respondent so that it can be correlated with his responses. This may include duty MOS, special army training, combat experience, etc. If it is really needed, it is usually asked for on the front page along with name.

b. If a questionnaire has over two pages, numeric page numbers should be used. They are ordinarily put at the center bottom of each page.

c. A questionnaire should not be crowded or cluttered in appearance. If it is, certain items might be missed.

d. Each item in a questionnaire should be numbered or lettered so it can be identified and referred to.

e. Sufficient room should be left for the respondent to write in his answers to open-ended questions.

f. Directions should be well displayed and unmistakably clear.

g. It is usually perferable to print the questionnaire in booklet form on both sides of the page, rather than have it duplicated on one side on the page and corner-stapled.

h. There is research evidence that an attractive questionnaire increases response rates.

i. Different colored pages or questionnaire forms may aid in the sorting of data and may have appeal to the respondents.

E. Use of Answer Sheets

As noted in Section IX-D 4, when possible, questionnaires should be designed to facilitate data collection, processing, and analyses. Hence, if the number of questions warrant it, consideration should be given to the use of separate answer sheets. An answer sheet can be designed for either hand or machine processing. A number of standard machine processable answer sheets are available, and copies will be included in a subsequent updating of this manual.

When considering the possible use of answer sheets, the following points should be kept in mind:

1. The use of a separate answer sheet may require a different set of abilities than responding on the questionnaire itself.
2. Depending upon their prior experiences with them, respondents may find it more difficult to use a separate answer sheet than to respond on the questionnaire sheet.
3. It is normally more difficult and time consuming for the respondent to use a separate answer sheet. (However, separate answer sheets have been used successfully for some purposes with fourth grade children).
4. When separate answer sheets are employed, the questionnaire booklets are reusable.
5. Respondents sometimes err in using the last spaces on a multiple choice answer sheet when there are more spaces than response alternatives. This can be avoided by the use of tailor-made sheets.

Chapter X: Considerations Related to Questionnaire Administration

A. Overview

Considerations related to the administration of questionnaires are discussed in this chapter, since such matters are obviously of concern when questionnaires are constructed. Questionnaire instructions are discussed in Section X-B, anonymity for respondents in Section X-C, motivational factors related to questionnaire administration in Section X-D. Administration time, characteristics of administrators, and administrative conditions are the topics of Section X-E, X-F, and X-G, respectively. The training of raters and other evaluators is the concern of Section X-H, while other factors related to questionnaire administration are considered in Section X-I.

B. Instructions

Care must be exercised in preparing instructions for questionnaires since they are quite likely to affect the way the respondent answers the questions. For example, even mildly anger arousing printed instructions may elicit responses of negativism.

Although further research is needed to fully determine the influence of instructions on responses, some practical guidelines can be offered:

1. It is sometimes preferred that an oral statement of questionnaire purpose be given to respondents. If this is not practical or a person with appropriate credibility and/or status cannot be supplied to make the statements, then a printed statement must suffice. (See Section IX-D 3 regarding printed introductions.)

2. Lengthy instructions for completing questionnaires should be avoided. They may tend to confuse the respondent rather than help him.

3. The option of orally presenting instructions is often available. When oral instructions are given they are usually given just prior to administering the questionnaire.

4. If instructions are given orally and an illustration is needed, a visual display should be available which may include a printed version of more complex instructions.

5. When questionnaires are group administered, it should be announced that aides will check each respondent's questionniare for completeness, if such a process can be implemented.

6. "Cute" examples on instructions should not be used. They will damage rapport and detract from the seriousness of the questionnaires, particularly for more mature and older respondents. It is best to use a neutral example that will be suitable for all respondents.

7. Obviously, instructions should be given in a way that all respondents can understand them. Care should be exercised about the level of vocabulary used.

An example is given on the following page of the instructions that might precede the items of a questionnaire. In this example the responses were to be given on a separate "answer" or response sheet.

TRAINING ATTITUDE QUESTIONNAIRE (BASIC AND AIT)

INSTRUCTIONS: The purpose of this questionnaire is to obtain information from you regarding training, working and living while in the Army's Basic Training and Advanced Individual Training (AIT) program. Your answers will help the Army to determine what conditions are in need of improvement, and will assist the Army in determining the actions they must take to improve training and the quality of life for new soldiers in the Army. Your honest opinions are, therefore, essential.

We have no need to know who you are personally. No effort will be made to identify either you or your unit. DO NOT WRITE YOUR NAME, SOCIAL SECURITY NUMBER, OR UNIT on either the questionnaire or the answer sheet.

Each question should be answered by circling the letter on your answer sheet which is next to the answer which best describes your feelings. See sample question below:

SAMPLE QUESTION: 3. How old are you?

a. 17
b. 18
c. 19
d. 20
e. 21 or older

If you are 19 years old, you should circle the letter c on your answer sheet for question 3, as has been done below, since the letter c corresponds to your correct age of 19 on the questionnaire.

QUESTION NUMBER	RESPONSES (CIRCLE ONE)				
01	a	b	c	d	e
02	a	b	c	d	e
03	a	b	[✓ c]	d	e
04	a	b	c	d	e

If you have any questions, please ask the questionnaire administrator for assistance. You will have 30 minutes to complete the questionnaire. We will all turn in our answer sheets and leave at the same time. Do not turn the page and start to work until instructed to do so.

C. Anonymity for Respondents

1. Factors to be Considered

There are several factors to be considered when deciding whether to require the respondent's name or other identifying information on a questionnaire. Some of the factors are supported by research, while others are not.

a. If the respondent supplied his name, he is aware that he can be identified and called back. If respondents do not have to give their names or similar information, most will believe that they cannot be identified and called back for any type of accounting after their questionnaires have been collected.

b. The perception of anonymity seems to depend not only upon whether a respondent gives his name, but also on the conditions under which the questionnaires are administered. For example, paper-and-pencil questionnaires are more anonymous than structured interviews.

c. The effects of anonymity seem to be related to the content of the questionnaire. This is particularly true when information on sensitive areas is collected. For general attitudes, it may not matter.

d. The effects of anonymity may also depend upon who administers the questionnaire, and the circumstances under which it is administered. Responses may be distorted when respondents are identified and under high threat.

e. Respondents may be more lenient when rating other personnel if they think they will be identified.

2. Implications of the Privacy Act of 1974

If the experimenter, test officer, or questionnaire writer desires to obtain certain types of personal information from a respondent, the federal Privacy Act of 1974, in turn, requires that certain information first be given to the candidate respondent. One may use DA Form 4368-R, 1 May 75 for the purpose of communicating this information to the respondent. The form is shown filled out on page X-C 3. In this particular example the research questions dealt with attitudes toward their treatment in the Army.

A second example, Figure X-C-1, illustrates a more compact format. The same elements of information called for by DA Form 4368-R have been communicated; it's just that that form was not used.

A privacy act statement is not necessarily required as a part of all questionnaires that are administered to Army personnel. It is not necessary where no personal information is being requested, and where the individual does not have to identify himself by name, SSAN, or other mark or characteristics. For example, no invasion of privacy is involved where soldiers are asked to anonymously evaluate some new/revised weapon, equipment, organization regarding effectiveness and/or acceptability.

DATA REQUIRED BY THE PRIVACY ACT OF 1974
(5 U.S.C. 552a)

TITLE OF FORM	PRESCRIBING DIRECTIVE AR 70-1

1. AUTHORITY

10 USC Sec 4503

2. PRINCIPAL PURPOSE(S)

The data collected with the attached form are to be used to research purposes only.

3. ROUTINE USES

This is an experimental personnel data collection form developed by the U.S. Army Research Institute for the Behavioral and Social Sciences pursuant to its research mission as prescribed in AR 70-1. When identifier (name or Social Security Number) are requested they are to be used for administrative and statistical control purposes only. Full confidentiality of the responses will be maintained in the processing of these data.

4. MANDATORY OR VOLUNTARY DISCLOSURE AND EFFECT ON INDIVIDUAL NOT PROVIDING INFORMATION

Your participation in this research is strictly voluntary. Individuals are encouraged to provide complete and accurate information in the interests of the research, but there will be no effect on individuals for not providing all or any part of the information. This notice may be detached from the rest of the form and retained by the individual if so desired.

FORM Privacy Act Statement - 26 Sep 75

DA Form 4368–R, 1 May 75

Figure X-C-1

A Second Example of a Privacy Act Statement

11B/C GRADUATE FIELD SURVEY
(Prescribing Directive: AR 600-46; TRADOC Ltr dtd 29 Aug 75)

INFORMATION PRIVACY ACT STATEMENT

1. Authority: 5 USC 301, 10 USC 3012, Authority for the Secretary of the Army to Issue AR's; 44 USC 3101, Authority for Collecting Necessary Data.

2. Principal Purpose: To collect data to evaluate the effectiveness of individual training received prior to joining one's initial unit of assignment.

3. Routine Uses: The data collected with this form are to be used for research purposes only. They will not become a part of any individual's record and will not be used in whole or in part in making any determination about an individual.

   The identifiers (name or Social Security Number) are to be used for administrative and statistical control purposes only. Full confidentiality of responses will be maintained in the processing of these data.

4. Mandatory or Voluntary Disclosure and Effect on Individual Not Providing Information: Voluntary - Your participation in this research is strictly voluntary. Individuals are encouraged to provide complete and accurate information in the interests of the research, but there will be no effect on individuals not providing all or any part of the information.

   This notice may be detached from the rest of this form and retained by the individual answering the questionnaire if so desired.

D. Motivational Factors

This section considers the effects of lack of motivation, and some ways of providing a desirable level of motivation to respondents during the questionnaire administration process.

1. Effects of Lack of Motivation

Generally, the results of any study will suffer distortion if those to whom the questionnaire is distributed are not sufficiently motivated. If they have the choice, they will not respond at all. If they do have to respond or are just minimally motivated, they may omit items, make patterned or random responses, or just generally respond poorly. As a result, the reliability and validity of the responses will be decreased and hence the results of the study left open to serious question.

2. Ego Involving Potential Respondents in the Study

There are a number of ways that motivation can be increased by ego involving potential respondents. Some of the ways are given below:

a. The special role of the respondent in the study can be emphasized.

b. Responsibility can be stressed when it is appropriate to do so.

c. The wording of cover letters, if used, affects ego involvement. Help may sometimes be requested on the basis of appealing to the self interests of the respondent. There is evidence that this type of appeal helps most with less educated respondents.

3. Stimulating the Return of Remotely Administered Questionnaires

Obviously, whatever ego involves potential respondents in a study also stimulates the return of remotely administered questionnaires, such as those distributed by mail. Other ways of stimulating the return or response rate are:

a. Return rates may often be significantly improved when a letter is sent in advance notifying the potential respondent that he will recieve a questionnaire and his help is needed in filling it out.

b. Stamped and addressed return envelopes can be sent with the questionnaire. There is evidence that this does increase response rate.

c. There is contradictory evidence about whether short questionnaires are returned more frequently than longer ones, but one would intuitively believe it to be true.

d. Followup reminders can be sent to those who do not promptly return their questionnaires. There is some question, however, regarding how much such followups increase response rate. At times it may not be cost effective, so maybe the decision should be a function of whether or not the initial return rate was adequate.

4. Use of Incentives

The evidence has been equivocal regarding the extent to which motivation is increased through the use of incentives. Incentives may include money, time off, special privileges, etc. Generally, however, it is agreed that incentives usually help increase the response rate with remotely administered questionnaires.

5. Other Motivational Factors Related to Questionnaire Administration

Many additional motivational factors related to questionnaire administration could be noted or inferred from other sections in this manual. Some of them are:

a. Respondents often have preferences for certain item formats, although sometimes such preferences do not seem to have an effect on results. Some subjects prefer rating scales to forced choice items. With forced choice some like the option of indicating the degree of applicability of each statement. Some do not like forced sort Q-sort (See Section IV-G.) Some prefer multiple category to two category options. These preferences may relate to familiarity of the respondent with given item types. There is not much that the questionniare designer can do about such preferences, except to note that they exist.

b. Motivation may be increased by offering feedback of study results to the respondent.

c. Every effort should be made to praise the respondents or potential respondents, to the extent that it is reasonable.

d. Long, vague, or boring questionnaire sessions should be avoided, since it will decrease respondent motivation.

e. Questionnaire administration sessions should not be scheduled when there are conflicts with other activities of greater interest to the respondents. Nor, in general, should they be schedules very early or very late in the day.

f. Volunteers are usually more motivated to fill out questionnaires than are nonvolunteers. However, their replies may be more biased.

g. When respondents are told that they may leave as soon as they have completed the questionnaire they usually do a much more hasty and unsatisfactory job than when they are given a specific time for completion, and are told that they cannot leave until the time period is up.

h. See Chapter XIV about the behavior of interviewers.

E. Administration Time

Little is known about the effects of questionnaire administration time on respondents motivation, or of the effects of setting time limits for completing questionnaires. The questionnaire administration period should generally have been determined in advance by pretesting. Although there will be some variability in the length of time taken to complete a questionnaire, there is remarkable consistency among those who are sincere in attempting to do an accurate and complete job of answering all questions.

When a questionnaire is administered to a group of respondents, the instruction should emphasize that all respondents will be given plenty of time to answer the questions. As indicated earlier in X-D 5 g, the instructions should not tell the respondents that they can leave as soon as they have finished the questionnaire, since many will then cut short their efforts to answer the questions. There is little hope of obtaining carefully considered evaluative responses on a questionnaire if the respondent knows that the faster he finishes the questionnaire the sooner he will be able to go home.

Questionnaire administration time is obviously related to questionnaire length, which is the topic of Section IX-C.

Every attempt should be made to determine the maximum time needed to complete a given questionnaire. If the questionnaire is group administered, the maximum time for the slowest respondents should usually be used in scheduling the administration of the questionnaire.

F. Characteristics of Administration

As with other areas of this manual, little has been established in the research literature about how the characteristics of questionnaire administrators affect the overall process with nonremotely administered questionnaires. The following items may be noted:

1. In most cases it is felt that the sex of the administrator has no effect on the responses received. There may, however, be certain motivational effects.

2. The military rank of the administrator may have an effect on the respondent, but no research has been performed to indicate this.

3. Any effect that the race of the administrator has on the respondent may be a function of the content material of the questionnaire e.g., race would be expected to influence responses on a race relations questionnaire more than on a questionnaire dealing with rifle comparisons. The effects should probably be viewed as the result of interaction between administrator and respondent characteristics, and the questions being asked.

4. See Chapter XIV about the influence on an interviewer on the interviewee.

G. Administration Conditions

Questionnaire administration conditions obviously cannot be controlled with remotely administered questionniares. With group administered questionnaires, the following guidance is offered:

1. Administration conditions should be provided which are most appropriate to the particular type of respondent completing the questionnaire.

2. Administration conditions have an effect on questionnaire responses. For example, different responses may be obtained if the questionnaire is filled out in a group situation on the job rather than individually at home.

3. When personnel are being rated, different ratings may be obtained depending on how acquainted the rater and ratee are.

4. For Army field test evaluations, the circumstances under which questionnaires must/can be administered will vary rather widely. There may be times when no writing surface(s) or pencils are available; clipboards and pencils should be supplied if this problem can be anticipated. If the needed materials cannot be brought to the respondents, then arrange to move them to a place where the materials and other environmental conditions are satisfactory.

5. Respondents should be required to give their answers without being influenced by other respondents. Achieving this requires respondents to be somewhat separated and/or to have the administrator(s) watching them. Simply instructing them not to consult with each other is usually not sifficient.

H. Training of Field Test Evaluators

An extended discussion of the training of raters and other test evaluators is not undertaken in the preliminary version of this manual. The following suggestions, however, can be offered about the general training of the Army field test evaluators. See Section X-B regarding questionnaire administration instructions.

1. Impress on test evaluators that they are supposed to answer the questionnaire based upon what they observe in the test. Stress the need for evaluations based only upon what was seen during the test exercise, regardless of any personal feelings or knowledge of concepts or equipment as might exist in a true combat environment (except in special instances where this is specifically asked for). To help identify and reduce prejudgment, a broad question might be included to permit the evaluator to express any bias he may have. It may be a question such as "Based on your personal experience, do you feel the "DPST" is a useful approach to real daily problems, i.e., outside a test exercise environment?" Such a question would permit the evaluator an outlet for preconceived opinions and attitudes which otherwise would color his view of the events observed during the exercise. On the other hand, in some situations the evaluator might feel it necessary to defend this personal judgment by biasing his answers to the remaining question answers!

2. Stress the importance of evaluators to the success of the test. Perhaps briefly indicate some actions which have been taken to implement concepts supported by evaluative data from previous tests.

3. Permit evaluators (particularly after the pilot test) to sound off about the forms and their perceived inadequacies, regardless of how unreasonable these complaints might be. The goal is to have all evaluators answering questionnaires understand that they are active contributors rather than just a means to an end.

4. Constantly examine completed questionnaires to insure that they have been filled out and understood. This procedure should continue throughout the entire series of tests.

5. Stress the notion that complete honesty and objectivity is needed. Sometimes evaluators try to please the test sponsors, to the detriment of the test.

6. Indicate to evaluators, perhaps on the top of all questionnaires or verbally, that they may make marginal note clarifications concerning their scale value selection for any rating question. This will increase posttest accuracy in determining questions which are scaled awkwardly or unclearly stated. This is particularly crucial during the pretesting or pilot test. Notes should be made regarding question structure immediately as they occur to the evaluator or the difficulty is likely to be forgotten.

7. Prior to having the evaluators complete questionnaires ask all, or a few randomly selected evaluators to verbally describe to the other evaluators what they believe each question is asking. This procedure will reduce differences between judges because of varying semantic interpretations. By the time of the actual exercise, all evaluators should generally agree, for example, on the meaning of "command and control effectiveness," "fire power potential," etc. If this is done, the criteria will have mutual acceptance.

8. Evaluators should be forwarned about biases such as the halo effect, central tendency, and others discussed in Chapter XII. If it is explained to the evaluator that these are common biases to which we are all subject, he will be better able to consider the fairness and accuracy of his observations.

9. The independent evaluation of each question should be stressed.

I. Other Factors Related to Questionnaire Administration

Some other factors related to questionnaire administration that have not been discussed in other sections of this manual are addressed below:

1. Respondents may at times be influenced by the title of the questionnaire. The word "test" should not be used in a title of a questionnaire at it may imply that it is a test of the respondent's knowledge.

2. A problem with Army field test evaluations concerns undue influence by the questionnaire administrator. It is sometimes necessary to use line officers from the units of the test subjects as questionnaire administrators. When outside administrators are used, they must be carefully instructed to make no comments whatsoever regarding their personal opinions of the items being evaluated. An offhand comment by a company commander administrator to his company regarding the "goodness" or "badness" of a piece of equipment or concept being evaluated can exert an influence sufficient to distort the results significantly from what they would otherwise have been.

3. The manner in which test subjects are selected and utilized in operational tests may affect the manner in which they respond to questionnaire items. For example, separate groups with no prior experience with either the test system or the current standard system could evaluate each system. This would exclude pretest biases, but test subjects would have no basis to compare the two systems. Alternatively, the same group of test subjects could use both systems in rotation. However, this procedure may result in a bias for or against one or both systems as a function of which was used first. In this respect too, personnel having extensive prior experience with a current standard system may introduce their pretest biases for or against that system when it is being evaluated against a candidate replacement system. The consequence of such considerations is that the type of system evaluation intended will govern the way evaluators and/or test subjects are selected and utilized. The methods of selection and utilization will influence the way questionnaires must be designed, and in turn suggest the types of problems likely to arise.

Chapter XI: Pretesting of Questionnaires

A. Overview

Even the most careful screening of a questionnaire by its developer or by questionnaire construction experts will usually not reveal all of its faults. Pretesting is an important and essential procedure to follow before administering any questionnaire. Its purpose is, of course, to find those overlooked problems and faults that would otherwise reduce the validity of the information obtained from the questionnaire responses. However, just any pretest will not do. One must know how to pretest the items and what to look for.

Some guidelines for pretesting questionnaires are given in this chapter. Pretesting may seem to some uninformed individuals to be a waste of time, especially when the author may have asked several people in his own office to critique the questions, or perhaps even asked a questionnaire specialist to critique it. However, pretesting is an investment that is well worthwhile. It is crucial if the decision that will result from the questionnaire is of any importance.

B. Guidlines for Pretesting Questionnaires

1. It is important that the respondents employed in pretesting be representative of the eventual target respondents. For example, if infantry enlisted men will perform in a test and then take the questionnaire, it should not be pretested with respondents who are armored officers; even infantry officers would not be satisfactory.

2. The pretest is more useful if it is conducted by someone who knows the operations to be performed in the test and who also knows the subject matter that the questionnaire covers. It is best if the question writer himself is knowledgeable about these operations and conducts the pretest.

3. Interview and pretest some of the pretest respondents one at a time. Ask each respondent to read each question and explain its meaning. Also ask him to explain the meaning of the response alternatives and to make his choice, and then ask him to explain why he made his particular choice. The respondents' answers will frequently reveal incorrect assumptions and possible rationales that the question writer never dreamed possible. They will also help to identify lack of understanding of particular words, vague or ambiguous phrases, ill defined or loaded questions, etc.

4. One good technique for pretesting is to have the respondent read each question aloud and then to tell you what it means. Any difficulties at all should be a cause for concern and revision.

5. During pretesting the respondents should be encouraged to make marginal notes on the questionnaire regarding sentence structure, unclear questions or statements, etc.

6. When attitude questions, especially, are being pretested, individuals who may hold minority views should be included. This will help identify loaded questions.

7. Open-ended questions may, and often should, be included in early pretest versions of a questionnaire in order to identify requirements for additional questions. Pretesting may also provide information that can be used to convert open-ended questions to multiple choice questions to facilitate data reduction and analysis.

8. Pretests for the selection of verbal anchors are valuable in building rating scale content validity and reliability. Rather than employing anchors which seem appropriate, the anchors used in the final scales should be selected as a result of analyses of pretests of respondents similar to those who will be participating in the final test.

9. While pretesting a questionnaire, a high proportion of respondents giving no response or a "Don't know" response should be a cause for concern. However, a low number of "Don't know" responses (especially for multiple choice items) does not guarantee that the question is good.

10. Often more than one pretest is needed. At times questionnaires may have to go through six or more pretests and revisions.

11. After pretesting, each question should be reviewed and its inclusion in the questionnaire justified. Questions that do not add significant information or that largely duplicate other questions can profitably be eliminated.

Chapter XII: Characteristics of Respondents That Influence Questionnaire Results

A. Overview

This chapter discusses some characteristics of respondents that influence questionnaire results. It therefore identifies some of the principal sources of error in the reporting of observations and/or the evaluation of performance in, for example, operational Army field tests. Additional research is required, however, to determine their relative contributions to error variance.

Sections XII-B and C, present a discussion of various biases, response sets, or other sources of error. There is some confusion in the literature regarding the use of these terms, but they are similar. A bias is: a tendency to deviate from a true value; a tendency to favor a certain position or conclusion; or an attitude either for or against a certain unproved hypothesis which prevents an individual from evaluating the evidence correctly. A response set or response bias refers to the tendency of a respondent to answer questions in a particular way almost independent of the content of the questions. And an error is simply a mistake or departure from correctness.

Section XII-D addresses the effects of attitudes of respondents on questionnaire results, while Section XII-E considers the effects of demographic characteristics on responses.

One of the main purposes of this chapter is to alert the questionnaire designer to some of the characteristics of respondents that influence questionnaire results. There are ways that some of the biases and errors can be controlled, but not all of them. And there appears to be no easy way of detecting the influence of a response set nor of neutralizing it. More detailed identification and control methods are areas of needed further research.

B. Social Desirability and Acquiescence Response Sets

Social desirability is a response set where persons answer according to the norms they believe society condones. It is the tendency to agree with items the respondent believes reflects socially desirable attitudes in order to show himself in a better light. Acquiescence response set is the tendency to consistently agree, to say "Yes," or to say "True." It is a general tendency to assent rather than dissent. Although there have been a number of studies about each, a detailed discussion of them is beyond the scope of this manual. Some comments about each are presented below.

1. Social Desirability Response Set

a. Social desirability response set seems to operate whenever the respondent has the opportunity to respond in terms of it. Some believe that its effect is so powerful that respondents would not tend to deviate from social norms in their answers even though their behavior denied what they said.

b. Several authors have identified respondents with a high social desirability response rate. They found these respondents to give more true responses to neutral items, to be more susceptible to social pressures, to more likely be introverts, and to score higher on a "lie" scale.

c. Faking or responding with socially desirable answers which are not true is part of the response set.

d. Anonymity fails to eliminate the social desirability response set.

e. The forced choice instrument format has been studied for its susceptibility to social desirability response set, a factor it was intended to control. Some authors found the forced choice method minimized the effects of social desirability, while others think the factor still needs additional control. One study concludes that in forced choice formats ambiguous items tend to be freer of social desirability response set than positively or negatively worded items. In any case, the evidence indicates that the social desirability problem is usually less in forced choice formats than in other item types.

f. Even card sorts need control to eliminate social desirability bias.

g. Procedures have been developed for controlling or balancing social desirability by using loaded items in the questionnaire and then adjusting the respondent's score. The social desirability score from the loaded items can also be correlated with each of the other items on the questionnaire. The responses on those items with a statistically significant correlation can then be corrected by moving the response one or more steps from the socially desirable response to give a more accurate result.

2. Acquiescence Response Set

   a. The acquiescence response set is defined as a behavioral attitude by the respondent to agree and accept, even if he must alter his original opinions to do so.

   b. The acquiescence response set seems to operate especially when statements are in the form of plausible generalities.

   c. The response set may occur more with difficult than with easy questionnaire material.

   d. Acquiescence response set may be a personality trait.

   e. There is a concern that social desirability and acquiescence response sets may be related in such a way that an individual with a tendency toward conformity will consistently reflect both biases.

   f. Controls for acquiescence response set have been researched. Stating the question stem in a netural manner may help minimize acquiescence. The effects of acquiencence response set may also be partially controlled by using two alternate questionnaire forms with the question stated positively on half of the forms and stated negatively on the other half. The balancing of scales (e.g., equal number of positive and negative points) may also be of value in counteracting acquiescence.

C. Other Response Sets or Errors

This section notes a number of other response sets or errors of which the questionnaire developer should be aware.

1. Error of Central Tendency

Some respondents tend to avoid endpoints on a scale, and pick a middle value regardless of their true feelings. It may be more common when the respondent is not very familiar with whatever he is being asked to rate. It may be counteracted by adjusting the strength of the response alternatives so that there are greater differences in meaning between alternatives near the ends of the scale than between alternatives near the center.

2. Extreme Response Set

On the other hand, some individuals tend to consistently select exaggerated choices for positions. It can be recognized when a respondent makes a pattern of answers which tend to be unevenly distributed toward one or both ends of a scale. Research indicates that this response set may be a personality characteristic.

3. Halo Effect

Halo effect was originally defined as a tendency, when one is estimating or rating a person with respect to a given trait, to be influenced by some other trait or by one's general impression of the person. It is, however, also applicable to ratings of other than people. For example, if a field test evaluator knows that a particular weapon system did well in one phase of a test, he may be influenced to give high ratings to the system in later test phases - even when the system performs poorly.

Most studies of ways to control halo effect have dealt with ratings of traits of personnel by other personnel, a matter not of great concern in this manual. The forced choice technique minimizes halo effect in some situations. Ratings will also be less distored if questionnaire items are constructed so as to relate to clearly observable aspects of behavior which do not overlap. It is doubtful that the influence of halo effects can be completely eliminated from the responses to any questionnaire.

4. Leniency Error

Leniency error refers to a general, constant tendency for a rater to rate either too high or too low in the direction of being too generous. It appears similar to halo effect except that it is independent of the trait or factor being rated. Some raters have an opposite tendency to rate too severly. In large groups of raters the opposite tendencies should balance out.

5. Logical Error

Logical error is also similar to halo effect. It is due to the fact that raters are likely to give similar ratings to traits or items that seem logically related to them. For example, a field test evaluator may know that a counterattack was extremely successful; he may therefore, reason that command and control was also very effective and should receive an equivalent high evaluation because a successful counterattack is a function of good command and control. Such reasoning assumes a dependence which may or may not be true. Logical error may be avoided in part by asking for judgments of objectively observable actions or behavior.

6. Proximity Error

Proximity error occurs when, due to the ordering of questionnaire items, the answer to one item results in an answer to a subsequent question being substantially changed from what it would otherwise have been. Little is known about its influence in field test situations; most research in this area has concerned the rating of personality trait variables.

7. Contrast Error

Contrast error refers to a tendency for a rater to rate others in the opposite direction from himself in regard to a trait. Little research has been done on this source of error.

8. Feedback Bias

Research shows that if observers are informed of experimental hypotheses and if they receive daily feedback indicating how well their data support the hypotheses, they will tend to report data supporting those hypotheses - even when the reverse is true! This bias does not seem to occur, however, when observers are informed only of the experimental hypotheses with no follow-up. Taking precautions to assure high levels of observer accuracy minimizes the bias.

D. Effects of General Pretest Attitudes of Respondents

Limited research has been conducted upon how the attitudes of a respondent influence questionnaire results. The following, however, should be noted:

1. Respondents at times base their ratings not on what is observed but on what they believed prior to the observation. Beliefs and opinions may affect results.

2. It is generally believed that judges used as part of the process of determining scale values can rate items without being influenced by their own attitudes. There is also some evidence to the contrary.

3. Unstable or changing responses to questionnaires may be caused by shifts in the mood of the respondent, relative values among the possible choices, and the degree of interest present in the question.

4. As questions become more ambiguous, responses normally become more attitudinally based.

5. It may be desirable to revise a questionnaire when norms of groups differ greatly from those with whom the questionnaire was pretested or previously administered.

E. Effects of Demographic Characteristics on Responses

Demographic characteristics have been shown to influence questionnaire results. Similarities of such variables among respondents often tend to be related to a response pattern. These variables include: age, religion, sex, intelligence, marital status, parenthood, socioeconomic class, nationality, urban or rural residence, income, rank and experience. Questionnaires should, therefore, be designed with the respondents background in mind. When there is a suspicion that demographic characteristics may affect response, the data should be analyzed by type of respondent.

Chapter XIII: Evaluating Questionnaire Results

A. Overview

An extended discussion on evaluating questionnaire results is currently outside the scope of this manual on questionnaire development. There are, however, some factors relating to the evaluation of questionnaire results that should be noted since they may influence how questionnaires are designed and developed. Section XIII-B considers the scoring of questionnaire responses, and Section XIII-C contains some notes about data analyses.

B. Scoring Questionnaire Responses

1. Practical Considerations

a. Both time and money can be saved by planning the questionnaire in line with scoring and tabulation requirements. The phrasing of questions and their sequencing and layout affect tabulation time.

b. A decision should be made ahead of time regarding whether the data will be tabulated by hand or machine.

c. Response alternatives should be precoded whenever possible.

d. Since it does not seem to matter if items are scrambled or in blocks according to content, blocking may be preferred due to greater hand scoring ease.

e. See Section IX-E regarding the use of answer sheets.

2. Other Considerations

a. There may be a justification for scoring rating scale items dichotomously according to the direction of response. It is sometimes done when bipolar scales are analyzed in terms of the proportion of responses in either direction of the basic dichotomy. The justification is based upon results that seem to indicate that composite scores reflect primarily the direction of responses and only to a minor extent their intensities.

b. One investigator found that many Likert-type rating scales consisting of 2 through 19 steps may be collapsed into two or three measurement categories for analysis with no lack of precision.

c. When working with paired comparison items with a "No preference" option, the "No preference" responses can often be either divided proportionate to the preference responses, or disregarded altogether. The basis for this suggestion is that respondents who claim neutrality appear to exhibit the same preference patterns as those who express a preference.

d. By using any one of several methods of scoring or transforming self-rating scale raw scores, it is usually possible to appriximate dyadic forced choice results with considerable saving in administration time, and a small gain in test-retest reliability.

e. The concurrent validity of questionnaires may be somewhat increased by using item weights obtained by expert scaling instead of conventional unit weights, but it may not be worth the effort.

f. Investigators sometimes use intensity scores as well as rating scale content scores. One way of obtaining an intensity score is to follow each question with the query "How strongly do you feel about this?" A second way involves weighting extreme responses (positive and negative) as 2, moderate responses as 1, and neutral responses as 0. These weights can then be summed for an intensity score.

C. Data Analyses

A detailed discussion of data analysis is beyond the scope of this manual; however, some basic data analysis issues have been mentioned in related chapters. Additionally, the following points are also noted:

1. Analyses of questionnaire responses is chiefly of two types: summary tabulations and statistical analyses. Tabulations are used primarily for the presentation of results. Statistical tests are used to determine whether the differences in the results are significant. Statistical literature is available which presents numerous tests usable in such analyses.

2. As part of the questionnaire development process, tentative (dummy) analysis tables should be developed to assure that the data to be obtained are appropriate.

3. Four kinds of measurement scales have been identified: nominal, ordinal, interval, and ratio. Appropriate statistical analyses are associated with each. Hence, the data analysis limitations of various forms of questionnaires should be considered before an instrument is designed. For example, less can be done statistically with open-ended questions than with ranking questions.

Chapter XIV: Interview Considerations

A. Overview

If properly used, the interview is an effective means of obtaining data. It is a technique in which an individual is questioned by a skilled and trained interviewer who records all replies, preferably verbatim in most cases. Most of the principals of questionnaire construction discussed in previous chapters pertain to the interview as well. This chapter, however, notes some issues specifically related to interviews.

Section XIV-B presents the distinction between structured and unstructured interviews. Interviewer's characteristics relative to the interviewee are noted in Section XIV-C. Situational factors are noted in Section XIV-D, while the topics of Sections XIV-E, F, and G are, respectively, training interviewers, data recording and reduction, and special problems. There is, unfortunately, little that can be recommended to avoid some of the problems noted in this chapter. The questionnaire developer should, in any case, be aware of them.

B. Structured and Unstructured Interviews

The term "structured" when applied to interviews is intended to emphasize that the interviewer employs a script of all the questions to be asked. In the unstructured interview the interviewer may know many of the topics to be covered but needs to learn more about the subject overall, so he is willing to be led by the interviewee even into digressions. Unstructured interviews may occur as a preliminary to preparing either a questionnaire or a structured interview script. One could use a questionnaire as the script for a structured interview if he already had the questionnaire developed, but not enough time to convert it to a more convenient format. The main difference between the structured interview and questionnaire is procedural.

The degree of proficiency required of interviewers in conducting an unstructured interview is generally not available during Army field test evaluations. A structured interview requires the interviewer to have only moderate skill and proficiency, and hence is usually preferred. The advantages of the structured interview include: the opportunity to probe for all the facts when the respondent gives only a partial or incomplete response; a chance to insure that the question is thoroughly understood by the respondent; and an opportunity to pursue other problem areas which may arise during an interview. The strucutred interview is almost always preferable to a questionnaire when the test group is small (10 to 20), and when time and test conditions permit.

As noted in Section II-B, unstructured interviews are not included within the definition of questionnaire used in this manual. They are, therefore, not discussed further.

C. Interviewer's Characteristics Relative to Interviewee

More research is needed to identify how characteristics of an interviewer affect the respondent. Some areas of concern are presented below.

1. Rank, Grade or Status of the Interviewer

For Army field test evaluations it is recommended that the interviewer should be of similar rank or grade to the individuals being interviewed. A difference in rank or grade introduces a bias in the data which has been found to substantially influence test results. Interviewees tend to give the answer they perceive the higher ranking interviewer favors. When the interviewer is of lower grade, the interviewee may not show respect and may not cooperate.

Evidence indicates that the greater the disparity between the status of the interviewer and that of the respondent, the greater the tendency for biased responses. The respondent tends to answer favorably in the eyes of the more serious interviewer.

Data suggest that in the interview situation the respondent tends to support the norms adhered to by the interviewer. Lower socioeconomic respondents may defer to the norms represented by a higher status interviewer. The effect, however, is related to the types of questions asked. Sensitive issues involving socially accepted or rejected answers will effect more bias.

2. Sex of the Interviewer

Differences in response patterns according to the interviewer's sex depend on subject matter as well as on the composition of the respondent populations and other characteristics of the specific survey situation.

3. Race of the Interviewer

The effects of the race of the interviewer on the respondent should probably be viewed as the result of interaction between interviewer and respondent characteristics. Respondents often give socially desirable answers to interviewers whose race differs from theirs, particularly if the interviewee's social status is lower than that of the interviewer and the topic of the question is threatening.

However, an interviewer's race can probaly establish different frames of reference even in nonsensitive areas. Particularly in regard to social issues, more valid results can be expected when the interviewer is of the same race as the respondent.

4. Experience of the Interviewer

It has been reported that there may be no significant differences between interview completion rates for experienced and inexperienced interviewers, and that the training and experience of the interviewer has no effect on the number of deviations they made from the instructions. However, regarding quality of interviews, all interviewers improve with experience.

D. Situational Factors

Among the situational factors that should be considered when interviews are used are the following:

1. It helps greatly if the interviewee perceives the interviewer as interested in hearing his comments, as willing to listen, and (if the situation requires) as willing to protect him from recrimination for being adverse in his evaluations.

2. Interviews should be conducted in a quite, temperature controlled environment where the respondent can be comfortable and relaxed. Each respondent should be interviewed in private, separate and apart from all others so that no other person hears or is biased by his responses.

3. The reinforcing behaviors of the interviewer have an influence on the responses collected, and at times may cause a respondent to change his preferences. Such comments as "good" or "fine" and such actions as smiling and nodding can have a decided effect on test results. Praised respondents normally offer more answers than unpraised ones. Praising respondents may also tend to reduce "Don't know" answers without increasing insincere or dishonest responses.

4. Interested respondents seem to be more subject to interviewer effects than uninterested ones.

E. Training Interviewers

Generally, interviewers require a certain amount of training. Such a discussion, however, is outside the scope of the initial version of this manual. Army personnel may check with the Army Research Institute-Field Unit closest to them for help in this area.

F. Data Recording and Reduction

In the structured interview both questions and answers are orally communicated. The interviewer may encode the answers on paper, or tape record the responses for later encoding (but only if the interviewee agrees to the taping and does not seem influenced by the presence of a recording device).

Other topics related to interview data recording and reduction are outside the scope of the initial version of this manual.

G. Special Interviewer Problems

This section notes some special problems related to interviews.

When interviews are used, the qualified interviewer will avoid leading, pressuring, or influencing the direction of an interviewee's evaluations. If a potential interviewer has strong preferences regarding the system(s) being tested, he should probably be disqualified.

Many studies have been conducted that show other biasing effects on the interviewer. Factors leading to significant effects of the interviewer upon results include: relatively high ambiguity in the concept of wording of the inquiry; the interviewer "resistance" to a given question; and additional questioning or probing. Interviewer bias can exist without being apparent, and the direction of bias is not necessarily uniform. The least interviewer bias is probably found with questions that can be answered "Yes" or "No." The bias can result from differences in interviewing methods, differences in the degree of success in eliciting factual information, and differences in classifying the respondent's answers. An interviewer's expectations may have a more powerful effect on the results than his ideological preferences.

Some interviewers have a tendency not to transmit printed instructions word for word. Hence total phrases may be eliminated and key words originally intended to focus the respondent's attention on some specific point are omitted or changed. Key ideas are lost, mainly through omission. Variability of interviewer performance seems to vary both across interviewers and within individuals.

An interviewer's attitude toward a question can communicate itself sufficiently to the respondent so that the meaning of the question is altered. Hence the nature of the survey and the survey organization are determining factors in whether or not the interviewer must follow the interview schedule verbatim, or may vary the wording.

Army Project Number
2Q763731A775

TCATA
DAHC19-74-C-0032

QUESTIONNAIRE CONSTRUCTION MANUAL ANNEX

Dr. Robert F. Dyer
Josephine J. Matthews
Josef F. Stulac
Dr. Calvin E. Wright
Dr. Kenneth Yudowitch
Operations Research Associates

Submitted by:
George M. Gividen, Chief
Fort Hood Field Unit

July 1976

Approved by:

Joseph Zeidner, Director
Organizations and Systems
Research Laboratory

J. E. Uhlaner, Technical Director
U.S. Army Research Institute for
the Behavioral and Social Sciences